INSIGHTS IN PLANT SYSTEMATICS AND EVOLUTION: 2021

EDITED BY: Jim Leebens-Mack and Gerald Matthias Schneeweiss
PUBLISHED IN: Frontiers in Plant Science

frontiers Research Topics

frontiers

Frontiers eBook Copyright Statement

The copyright in the text of individual articles in this eBook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this eBook is the property of Frontiers.

Each article within this eBook, and the eBook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this eBook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or eBook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-88976-750-2
DOI 10.3389/978-2-88976-750-2

About Frontiers

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

Frontiers Journal Series

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

Dedication to Quality

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

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

INSIGHTS IN PLANT SYSTEMATICS AND EVOLUTION: 2021

Topic Editors:
Jim Leebens-Mack, University of Georgia, United States
Gerald Matthias Schneeweiss, University of Vienna, Austria

Citation: Leebens-Mack, J., Schneeweiss, G. M., eds. (2022). Insights in Plant Systematics and Evolution: 2021. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88976-750-2

Table of Contents

frontiers in Plant Science

REVIEW
published: 16 February 2022
doi: 10.3389/fpls.2022.807302

Why Do Heterosporous Plants Have So Few Chromosomes?

Sylvia P. Kinosian[1], Carol A. Rowe[2] and Paul G. Wolf[3]*

[1] Negaunee Institute for Plant Conservation Science, Chicago Botanic Garden, Glencoe, IL, United States, [2] Earth System Science Center, The University of Alabama in Huntsville, Huntsville, AL, United States, [3] Department of Biological Sciences, The University of Alabama in Huntsville, Huntsville, AL, United States

The mechanisms controlling chromosome number, size, and shape, and the relationship of these traits to genome size, remain some of the least understood aspects of genome evolution. Across vascular plants, there is a striking disparity in chromosome number between homosporous and heterosporous lineages. Homosporous plants (comprising most ferns and some lycophytes) have high chromosome numbers compared to heterosporous lineages (some ferns and lycophytes and all seed plants). Many studies have investigated why homosporous plants have so many chromosomes. However, homospory is the ancestral condition from which heterospory has been derived several times. Following this phylogenetic perspective, a more appropriate question to ask is why heterosporous plants have so few chromosomes. Here, we review life history differences between heterosporous and homosporous plants, previous work on chromosome number and genome size in each lineage, known mechanisms of genome downsizing and chromosomal rearrangements, and conclude with future prospects for comparative research.

Keywords: ferns, homospory, genome evolution, meiosis, heterospory, chromosome evolution

OPEN ACCESS

Edited by:
Gerald Matthias Schneeweiss, University of Vienna, Austria

Reviewed by:
Eva Maria Temsch, University of Vienna, Austria
Petr Bures, Masaryk University, Czechia
Martin Burd, Monash University, Australia

****Correspondence:***
Sylvia P. Kinosian
sylvia.kinosian@gmail.com

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

Received: *02 November 2021*
Accepted: *26 January 2022*
Published: *16 February 2022*

Citation:
Kinosian SP, Rowe CA and Wolf PG (2022) Why Do Heterosporous Plants Have So Few Chromosomes? Front. Plant Sci. 13:807302. doi: 10.3389/fpls.2022.807302

INTRODUCTION

The nuclear genetic material of eukaryotes is contained within chromosomes. The number, length, and centromere location of chromosomes in an organism (the karyotype) varies considerably among lineages (Ohno, 1984; Schubert and Lysak, 2011). In plants, chromosome number is often phylogenetically conserved (Manton, 1950; Wagner and Wagner, 1979; Weiss-Schneeweiss and Schneeweiss, 2013). Changes in chromosome number or structure can alter the balanced chromosome pairing that is critical for cell division, leading to sexual sterility or death. Thus, an understanding of how plant chromosome numbers evolve has implications for plant reproductive biology, systematics, and genome evolution, among other processes (Haufler, 2002; Li Z. et al., 2020; Fujiwara et al., 2021).

In vascular plants, there is a striking disparity in chromosome number between homosporous and heterosporous lineages (Klekowski and Baker, 1966): homosporous lineages have high chromosome numbers and often larger genomes, compared to heterosporous ones (Leitch and Leitch, 2013; **Figure 1**). Thus, several studies have asked why homosporous plants have so many chromosomes (Klekowski and Baker, 1966; Klekowski, 1969; Nakazato et al., 2008;

Barker and Wolf, 2010). However, homospory is the ancestral condition from which heterospory has been derived several times (Bateman and DiMichele, 1994; **Figure 2**). Given that the character state of high chromosome numbers in homosporous plants is ancestral (Clark et al., 2016; Carta et al., 2020), it makes more sense evolutionarily to ask why heterosporous plants have comparably so few chromosomes.

This review covers the life history differences between heterosporous and homosporous vascular plants, previous work on chromosome number and genome size in each lineage, and known mechanisms of genomic and chromosomal change. We address the current evidence for the processes controlling genome evolution and the variation in genome downsizing rates between homosporous and heterosporous plants, as well as how the pattern of spore production might be related to these mechanisms. We conclude with prospects for research on this relationship and what types of data will be needed to solve a mystery that has haunted botanists for over half a century.

LIFE HISTORY OF HOMOSPOROUS AND HETEROSPOROUS VASCULAR PLANTS

All seed plants, and some spore-dispersed plants, are heterosporous (**Figure 2**). They produce two different types of sporangia, resulting in the small microspores (sperm-producing) that develop to become microgametophytes and larger megaspores (egg-producing) that develop to form the megagametophytes. In seed plant microsporogenesis, all four meiotic products are retained and grow into pollen grains. In contrast, during megasporogenesis, only one of the four meiotic products survives. There are some exceptions to this process, including apomictic species where a polar body fertilizes the megaspore, as well as species where the polar bodies develop into endosperm (e.g., Schmerler and Wessel, 2011; Noyes and Givens, 2013).

Homosporous plants comprise most ferns, some lycophytes, as well as bryophytes (the latter are non-vascular). They produce a single type of spore that germinates into a gametophyte theoretically capable of producing both egg and sperm. Most ferns are homosporous, but one clade of aquatic ferns (Salviniales) has evolved to become heterosporous. Heterospory also evolved independently in the lycophytes. It is estimated that heterospory has evolved at least 11 times throughout the history of land plants (Bateman and DiMichele, 1994). However, we only have three independent lineages of extant heterosporous plants (**Figure 2**) with which to test patterns of associated characters. These extant heterosporous lineages provide us with natural replicates to study the evolution of this trait, as all three lineages have homosporous sister lineages for comparison.

HISTORY OF CHROMOSOME RESEARCH IN PLANTS

The study of plant chromosomes dates back to the nineteenth century when researchers began using microscopy to study cell biology. In the 1880s, the stages (Flemming, 1965) and timing (Strasburger, 1894) of cell division were first described, although the exact nature and importance of chromatin were not yet understood (Volkmann et al., 2012). In 1888, von Waldeyer-Hartz coined the term chromosome, and the structure and heritability of chromosomes became established soon after (Strasburger, 1894; Cremer and Cremer, 1988; Winkelmann, 2007). Around the turn of the twentieth century, researchers noticed plants with doubled numbers of chromosomes (e.g., *Oenothera,* Lutz, 1907; Strasburger, 1910). This work resulted in the theory that genome doubling restored fertility to hybrid plants (Winge, 1917). During the following decades, additional work extended the knowledge base of plant cytology, with a wide range of heterosporous study systems including *Nicotiana* (Clausen and Goodspeed, 1925), *Oenothera* (Gates et al., 1929), *Viola* (Clausen, 1927), *Gossypium* (Skovsted, 1935), and the Salicaceae (Blackburn and Harrison, 1924).

Around the same time, work also began on homosporous fern genetics (Lang, 1923; Andersson-Kottö, 1927, 1929, 1938). These initial studies provided the first chromosome counts and crossing experiments in ferns; importantly, they presented the theory that ferns with high chromosome numbers had diploid, not polyploid, inheritance (Andersson-Kottö, 1929, 1938; Haufler, 2002). In the 1950 book, *Problems of cytology and evolution in the Pteridophyta* (Manton, 1950), the chromosome numbers of about 100 species of ferns were published for the first time. This work provided an important reference for fern cytology and helped establish the importance of base chromosome numbers in classification (Manton and Sledge, 1954).

In 1966, the first connection was made between the differences in chromosome numbers of homosporous and heterosporous ferns. Klekowski and Baker (1966) used previously published chromosome counts to show that within ferns and lycophytes, homosporous species had an average sporophytic count of $2n = 115$, whereas for heterosporous species it was $2n = 27.24$. In comparision, in angiosperms (all heterosporous) was $2n = 31.98$. These findings have been substantiated by modern methodologies. Additionally, researchers have been able to evaluate the chromosome number and genome size for the ancestors of angiosperms and spore-dispersed plants. The ancestral gametophytic chromosome number for angiosperms has been estimated to be $n = 7$ (Carta et al., 2020), whereas in homosporous ferns this is estimated to have been $n = 22$ (Clark et al., 2016). These findings indicate that chromosome number may be influenced by the shift in life history to heterospory, although the mechanisms are still unclear (Clark et al., 2016; Carta et al., 2020; Fujiwara et al., 2021; Szövényi et al., 2021).

We downloaded (in January 2021) chromosome counts from 377,715 records in the Chromosome Counts Database version 1.47 (Rice et al., 2015). In addition, we examined data on genome size from the Plant DNA *C*-values Database (Pellicer and Leitch, 2020). Here, we also analyze the data, removing likely recent polyploid species from the analysis to include only the base chromosome number for each species. We did this by allowing for aneuploid and dysploid

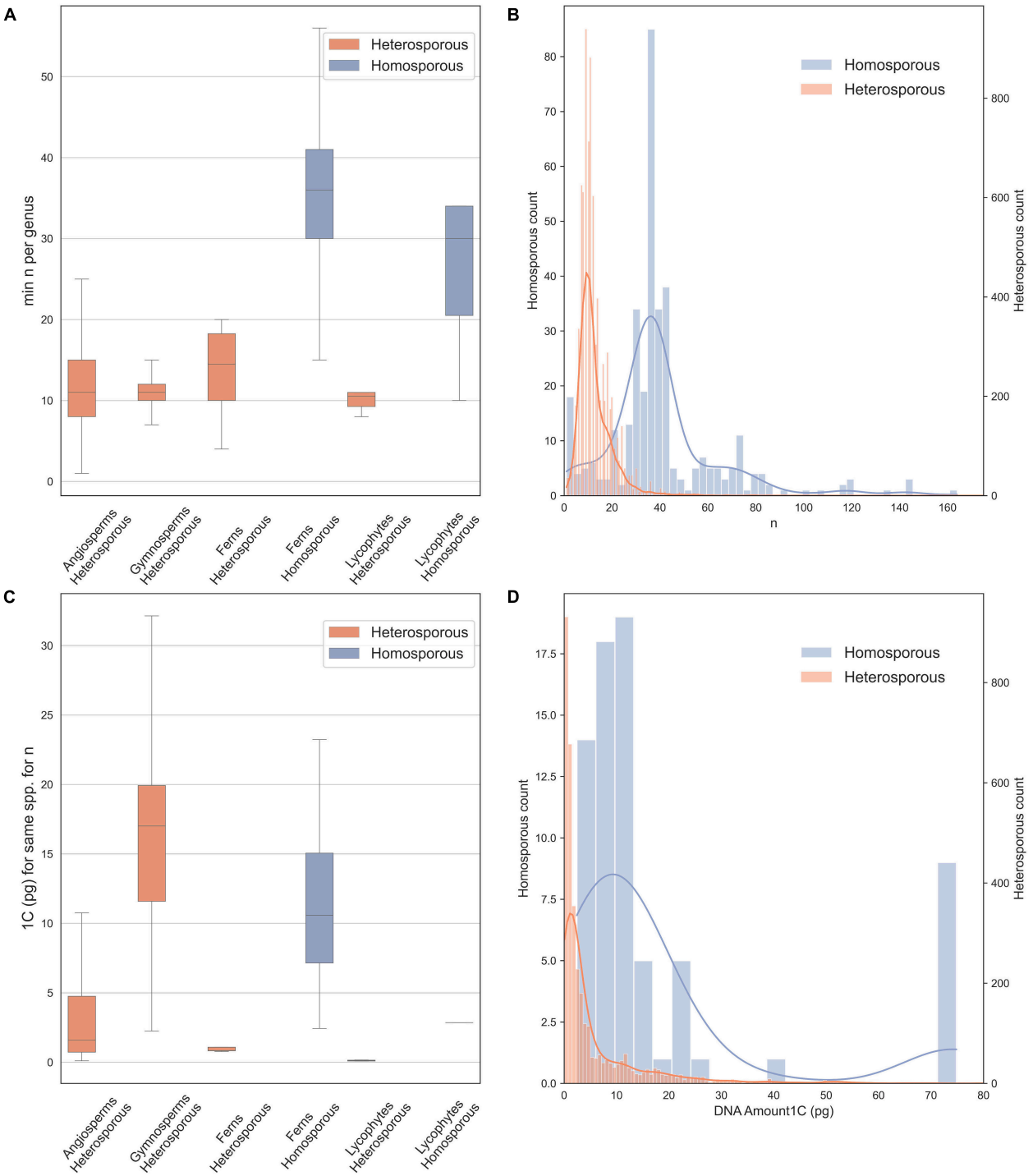

FIGURE 1 | Chromosome counts and genome sizes for heterosporous and homosporous plant clades. Data were downloaded in January 2021 from the Chromosome Counts Database version 1.47 (Rice et al., 2015) and the Plant DNA *C*-values Database (Pellicer and Leitch, 2020). **(A)** Boxplots of chromosome counts (n = 7,900), including only the minimum reported count for a genus, boxes are central two quartiles and whiskers are interquartile range × 1.5 (data from Rice et al., 2015). **(B)** Histogram with kernel density plot of chromosome counts for homosporous plants (blue, right y-axis; mean n = 42.25) and heterosporous plants (orange, left y-axis; mean n = 12.74)—note scales are different to account for unequal sample size. **(C)** Boxplots of genome size estimates (1C), included are all possible minimum values for each species. **(D)** Histogram with kernel density plot of genome size estimates for homosporous plants (blue, right y-axis; mean 1C = 12.52) and heterosporous plants (orange, left y-axis; mean 1C = 6.43), on different scales.

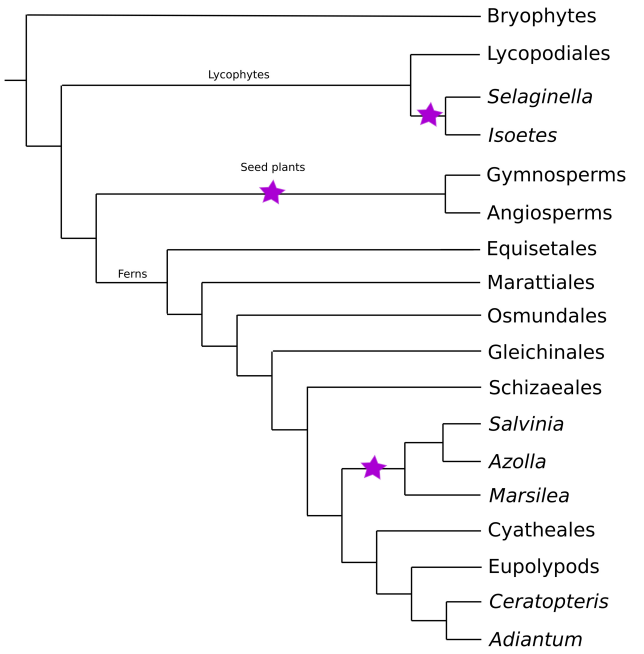

FIGURE 2 | Land plant phylogeny (PPG, 2016). The three major clades of vascular plants (lycophytes, seed plants, and ferns) are shown, along with the three heterosporous lineages denoted with purple stars. Tip labels include phylogenetically important clades as well as clades with current or forthcoming genomic resources.

change, and retaining species within a genus, with up to 1.2 times the minimum chromosome number recorded for the genus. We show chromosome numbers (**Figures 1A,B**) and genome sizes (**Figures 1C,D**) for homosporous and heterosporous vascular land plants. We present box plots for four heterosporous and two homosporous lineages (**Figures 1A,C**) and then also compare the distributions of all homosporous plants versus all heterosporous plants (**Figures 1B,D**). Note that our plots (**Figures 1B,D**) include histograms and kernel density because the sample sizes were very different between the groups and binning sizes made comparisons difficult.

On average, homosporous plants have more than three times as many chromosomes and more than three times the genome size as heterosporous plants (**Figure 1**). In all four heterosporous lineages (heterosporous lycophytes and ferns, gymnosperms, and angiosperms; **Figure 2**), there is a reduction in chromosome number. The pattern for genome size, however, is slightly more complex. Gymnosperms are anomalous with large genomes, despite relatively low chromosome numbers. It has been hypothesized that at least some gymnosperms have an unusually high density of long terminal repeat retrotransposons (LTR-RTs), responsible for their large genome sizes (Nystedt et al., 2013). Considering both chromosome number and genome size, there are differences in genome architecture and evolution between homosporous and heterosporous plants. What remains to be discovered are the genetic mechanisms that control these differences, and how they are influenced by life history.

HOW CAN THE PATTERN OF SPORE PRODUCTION BE RELATED TO CHROMOSOME NUMBER?

At first glance, it seems unlikely that aspects of chromosome number could be related to the spore types. Why would an evolutionary transition to heterospory accompany the derived character state of small genomes and low chromosome numbers? This question has challenged botanists for decades, and no simple hypothesis has yielded a satisfactory explanation. In the 1960s, starting with the observation of high chromosome numbers in homosporous pteridophytes (Klekowski and Baker, 1966), Klekowski (1969, 1972, 1973) developed a testable hypothesis for a causal relationship, based on the different reproductive modes that can occur in heterosporous versus homosporous plants. In heterosporous plants, reproduction can proceed *via* sporophytic selfing (sperm and egg from the same parent plant, but different gametophytes) or outcrossing (sperm and egg from different plants). In homosporous plants, both sporophytic selfing and outcrossing are possible, but an additional reproductive mode can occur called gametophytic selfing (Haufler et al., 2016). This extreme form of self-fertilization occurs when an egg is fertilized by a sperm from the same gametophyte. A gametophyte generates gametes *via* mitosis, so all gametes are genetically identical; therefore, a sporophyte created from gametophytic selfing will be homozygous at every locus in the genome.

Klekowski proposed that ferns primarily reproduced *via* gametophytic selfing and consequently would have high genetic load if they were diploid. Therefore, ferns with high chromosome numbers must have polyploid, not diploid, inheritance (Klekowski and Baker, 1966), directly opposing early work on fern genetics (Andersson-Kottö, 1929). Klekowski proposed that if these homosporous species became polyploid through hybridization (allopolyploidy) then diverse alleles could be maintained across the different parental genomes, reducing genetic load despite extreme inbreeding (Klekowski, 1976). This occurs *via* homoeologous heterozygosity: when matching chromosomes (homeologs) from different parental genomes in a hybrid carry distinct alleles (Glover et al., 2016). Furthermore, if the homoeologous chromosomes could pair, even a sporophyte that is homozygous at all homologous loci could create genetically variable meiotic products (Klekowski, 1972, 1973). Thus, the increased chromosome sets were needed to overcome the extreme mating system in homosporous plants.

Throughout the 1970s, researchers gathered data that seemed to support Klekowski's hypothesis. Evidence from chromosome studies, breeding studies, and genetic analyses seemed to indicate that homoeologous recombination was possible in homosporous ferns (Hickok and Klekowski, 1973; Klekowski and Hickok, 1974; Chapman et al., 1979; Hickok, 1979). Most of this work came from lab experiments, but starting in the 1980s, researchers began to examine how homosporous plants

behaved in natural settings. The first observation was that homosporous ferns appeared to be mostly outcrossing (Haufler and Soltis, 1984; Gastony and Gottlieb, 1985; Wolf et al., 1987), evidence that did not support high levels of gametophytic selfing which would be required to provide the selective pressures under Klekowski's hypothesis. Furthermore, studies applying electrophoresis of enzymes showed that even ferns with high chromosome numbers expressed the typical diploid number isozymes (Wolf et al., 1987). This was subsequently confirmed with genomic sequencing in the homosporous fern *Ceratopteris* (Marchant et al., 2019). Such evidence suggested that even species that appear to be polyploid behave genetically as diploids (Haufler and Soltis, 1986; Haufler, 1987, 1989) and that gametophytic selfing might actually be rare in homosporous ferns (Haufler et al., 2016).

Missing from this research, however, was a robust evolutionary framework for comparative analyses; such a phylogenetic perspective is critical for inferring the evolutionary processes that influence genome structure. Reconstructing the phylogeny for land plants started in the 1990s when *rbcL* was developed as a phylogenetic marker for angiosperms, followed shortly by large-scale phylogenetic analyses of ferns (Pryer et al., 1995). By early in the twenty-first century, there was a good working hypothesis for relationships among the major groups of ferns (Pryer et al., 2004). Researchers also began assembling plant genomic resources. Sequencing the *Arabidopsis* genome (The Arabidopsis Genome Initiative, 2000) was an important first step, soon followed by many more seed plant genomes. The 1000 Plant Transcriptomes project was a major step in generating genetic data for species across the land plant phylogeny (One Thousand Plant Transcriptomes Initiative, 2019). The first linkage map for a homosporous fern was published for *Ceratopteris richardii* (Nakazato et al., 2006). However, the first fern genomes were not completed until 2018 for heterosporous ferns (Li et al., 2018), and 2019 for homosporous *Ceratopteris* (Marchant et al., 2019). Whole genome sequences are also published for liverworts (Bowman et al., 2017), mosses (Rensing et al., 2008), and hornworts (Li F. W. et al., 2020). These bryophyte genomes provide us with an outgroup for all vascular land plants. The current phylogenetic, genomic, and transcriptomic resources are very close to providing the resources necessary to answer some of the old questions regarding heterosporous and homosporous genomes and chromosome evolution (Fujiwara et al., 2021; Szövényi et al., 2021).

MECHANISMS OF GENOMIC AND CHROMOSOMAL CHANGE

Despite limited phylogenetic, genomic, and transcriptomic resources, there is a growing body of research on broad patterns of chromosomal and genomic evolution between homosporous and heterosporous plant genomes. In most groups of organisms, there is not a good correlation of chromosome number with genome size. The reasons for this are complex and involve the, often rapid, loss of genetic material after a whole-genome duplication (WGD) event (Leitch and Bennett, 2004). In contrast, some studies suggest that genome size is positively correlated with chromosome number in homosporous ferns (Nakazato et al., 2008; Bainard et al., 2011; Clark et al., 2016; Fujiwara et al., 2021), indicating that ferns have fundamentally different mechanisms of genome downsizing compared to other organisms (Barker and Wolf, 2010; Leitch and Leitch, 2012). Here, we cover some hypotheses about the differences in genome downsizing, architecture, and chromosome structure between homosporous and heterosporous plants.

Genome Downsizing

In angiosperms, polyploidy (and associated WGD) has played a major role in the evolution of the vast majority of species, including both recent and ancient WGD events (Cui et al., 2006; Van de Peer et al., 2009, 2017). However, the long-term evolutionary effects of WGD are complex, and polyploid lineages include a mix of evolutionary dead ends as well as critical lineages that survive, exploit new niches, and radiate, perhaps as a consequence of polyploidy (Van de Peer et al., 2017). When polyploids are initially formed they have twice as many genes as they typically need, relaxing selection pressure on retention of the duplicated copies. This, combined with a documented breakdown in the meiotic process in polyploids (Ramsey and Schemske, 2002; Chester et al., 2012) can lead to rapid loss of chromosomal segments, resulting in downsizing of the genome (Leitch and Bennett, 2004; Li Z. et al., 2020; Bowers and Paterson, 2021). Studies of recent allotetraploids reveal extensive chromosomal variation, including intergenomic translocations, which all appear to be part of the process of rapid genome downsizing in angiosperms (Lim et al., 2008; Chester et al., 2012). These genomic changes ultimately lead to the restoration of disomic inheritance and bivalent chromosome pairing in lineages that have experienced WGD (Li Z. et al., 2020).

Much less is known about genomic change following WGD in homosporous plants, which has yet to be studied extensively (Szövényi et al., 2021). Overall, it is thought that genome downsizing proceeds slower in ferns than angiosperms (Barker and Wolf, 2010), and diploidization in ferns may be driven by pseudogenization and/or gene silencing rather than gene loss (Haufler, 1987; Barker, 2013; Clark et al., 2016). The latter process potentially leads to larger genomes, or rather genomes that do not decrease in size following polyploidization. Differences in downsizing rates may be a consequence of several processes. If fern genome downsizing is slower, it could be because ferns lose fewer chromosomal regions per generation. However, the cause could also be a reduction in the size, not the rate, of the chromosomal segments being lost. Recent work suggests that the rate of genome evolution may influence speciation rate. This is supported by high rates of genome evolution in the species-rich Polypodiales, which contain over three-quarters of extant fern diversity (PPG, 2016; Fujiwara et al., 2021). The mechanisms influencing fern genome evolution, however, are less clear. This illustrates the need for experiments that track chromosomal changes following polyploidy in ferns, and compare this to genome downsizing in heterosporous plants (Lim et al., 2008; Chester et al., 2012). Understanding the processes that influence genome architecture and the rate of genomic change

between lineages may help advance our knowledge of broad-scale speciation dynamics of all land plants (Fujiwara et al., 2021).

Genome and Chromosome Architecture

In addition to genome size, heterosporous and homosporous genomes have some critical differences in structure and composition. Our current knowledge of these differences in homosporous plants is limited to genome skimming and transcriptome studies, but these are important to inform future comparative genomic work. In heterosporous seed plants (angiosperms and gymnosperms), one aspect of genome size variation is transposable elements, such as LTR-RTs, which increase in copy number over time and can cause even diploid genomes to become very large (Wendel et al., 2016). Baniaga and Barker (2019) found that LTR-RT are also present in homosporous fern genomes, but have much older insertion times than in angiosperm taxa; this means that the LTR-RTs of homosporous ferns have had time to increase in copy number, inflating genome size. In addition, they found that heterosporous ferns (*Azolla* and *Salvinia*) and lycophytes (*Selaginella*) have LTR-RTs insertion times more similar to angiosperms than to homosporous ferns (Baniaga and Barker, 2019). These findings are consistent with the observation that homosporous fern genomes have a greater proportion of repeat elements than angiosperm genomes (Wolf et al., 2015). Baniaga and Barker (2019) hypothesized that LTR-RTs are associated with high methylation in homosporous ferns (Takuno et al., 2016): LTR-RTs are silenced by methylation, and methylation can also silence genes on the same chromosome because it targets repeat elements. Homosporous ferns may ultimately not purge these methylated and silenced LTR-RTs and other genes, leading to their large genome size (Baniaga and Barker, 2019). However, long-read sequencing and high-quality genome assemblies are needed to understand the evolutionary dynamics of repeat elements in homosporous ferns. Future comparative genomic work could investigate the role of mating systems and life history on repeat elements in homosporous and heterosporous genomes (Baniaga and Barker, 2019). Additionally, genomic data from a homosporous lycophyte (Lycopodiales) is needed to compare gene structure across homosporous lineages.

The distribution of genes and repeat elements along chromosomes also differs between heterosporous and homosporous plants. The structure of angiosperm chromosomes seems to be fairly well conserved, with most genes occurring between the repeat-rich pericentromeric and telomeric regions (Szövényi et al., 2021). Comparatively, in seed-free plants, genes and repeats are interspersed along the chromosomes rather than separated into repeat-rich and gene-rich areas (Banks et al., 2011; Liang et al., 2020). However, this has only been studied in mosses and lycophytes (Banks et al., 2011; Shakirov and Shippen, 2012; Bowman et al., 2017; Li F. W. et al., 2020). Chromosome-scale genome assemblies are needed to investigate such structure in homosporous ferns, as well as understand the variation in chromosome structure between homosporous lineages.

Variation in chromosome size is another major difference between heterosporous and homosporous plants, with a 3,100-fold variation in angiosperms compared to 31-fold in homosporous ferns and lycophytes (Clark et al., 2016). Our understanding of the mechanisms controlling chromosome size, particularly in ferns, is limited (Szövényi et al., 2021). Schubert and Oud (1997) showed that there is an upper limit to chromosome arm length in angiosperms, and Liu et al. (2019) found support for conservation of chromosome size within the fern genus *Asplenium*. Work in angiosperms has found that mitotic divisions may fail when the chromosome arm: spindle length ratio is above a certain point (Schubert and Oud, 1997). Investigating this ratio in homosporous and heterosporous ferns and lycophytes could be a fruitful avenue of study, as there may be fundamental differences in cell division between these lineages and seed plants. Overall, investigating patterns of both chromosome- and genome-scale structure will greatly benefit our understanding of both heterosporous and homosporous genomes.

DISCUSSION: COMPARATIVE ANALYSES OF HOMOSPORY AND HETEROSPORY

If we are to understand the relationship between genome architecture and spore type, then the necessary studies must begin with comparative analyses from a phylogenetic perspective. Given the three independent origins of heterospory (**Figure 2**), we ask what genomic characteristics are uniquely shared on the three branches subtending these heterosporous clades? Understanding *what* parts of the genome have changed following a transition to heterospory is critical for developing hypotheses explaining *why* these evolutionary steps have occurred. Within a few years, we should have sufficient numbers of homosporous genomes to search for statistically significant differences in gene family expansion and contraction, signatures of selection, trends in rates of pseudogenization, distribution and insertion rates of various groups of transposable elements and retrotransposons, and comparative patterns of synteny and general genome architecture. This work would then inform novel hypothesis-driven approaches that could include comparative analyses of genes related to meiosis and chromosome structure, such as those associated with spindle fiber genes, cell cycle genes, telomere structure, centromere structure, kinetochores, and recombination. We would also benefit from a return to chromosome analysis, but with genomic perspectives, using approaches such as fluorescent *in situ* hybridization combined with chromosome painting (e.g., Schubert and Lysak, 2011; Šimoníková et al., 2019).

We currently have the resources to explore some potential drivers of genomic change in homosporous ferns. To explain the observed chromosomal differences between homosporous and heterosporous genomes, a few theories have been proposed based on the disparate life histories of these two groups. For example, even if gametophytic selfing is not particularly common in homosporous ferns, it could still play a role in genome evolution, especially in chromosome structure. Marchant (2019) suggested that if there is a major chromosomal change in meiosis, the resulting gametophyte could successfully self-fertilize because both sperm and egg would carry the same

mutation. All chromosomes would pair without issue, fixing a new chromosome arrangement. Such an event could not occur in a heterosporous plant with obligate gametophytic outcrossing (Marchant, 2019). Gametophytic selfing would reduce negative selection on chromosomal malformations or other components such as repetitive elements (Baniaga and Barker, 2019) in homosporous plants, potentially leading to larger and more dynamic genomes than in heterosporous plants. Similarly, dysploidy and polyploidy might be more successful in lineages with gametophytic selfing, such as Ophioglossaceae, which have underground gametophytes (Soltis and Soltis, 1986; Hauk and Haufler, 1999), as well as very large genomes (Bainard et al., 2011).

This hypothesis proposed by Marchant (2019) suggests that homosporous fern genomes are more stable following WGD, due to gametophytic selfing. This would relax evolutionary pressures to downsize the genome, as all chromosomes could pair without issue, resulting in more consistent successful gamete production. One could test this hypothesis by examining natural and artificial polyploids to see if genomic segments are lost as fast as they are in seed plants (e.g., Lim et al., 2008; Chester et al., 2012). Furthermore, chromosome structural analyses will be needed to test Marchant's (2019) hypothesis for fixation of chromosomal changes *via* gametophytic selfing. This could be accomplished through comparative karyotype analysis or computational analysis of synteny in a series of sister taxa.

Another potential driver of chromosome evolution in plants is transmission ratio distortion (TRD): the preferential inheritance of one allele over the other (Huang et al., 2013). TRD can be caused by several mechanisms, including germline selection (Hastings, 1991) and meiotic drive (Pardo-Manuel et al., 2001). Both of these processes can influence genome structure by preferentially selecting for a certain gene or chromosome structure, although they affect homosporous and heterosporous plants differently. During megasporogenesis in most heterosporous angiosperms, four cells result from meiosis but only one of the two outer cells survives to become the megagametophyte; the remaining polar bodies die. Certain genes, chromosomes, or portions of chromosomes can be preferentially transported to these outer cells, increasing the probability of being passed to the next generation. There are physical attributes of chromosomes that help to transport all or a portion of a chromosome to the outer cells during meiosis (Burt and Trivers, 2009). For example, there is a bias against inversions and for deletions, because smaller chromosomes move faster along the spindle and therefore are more likely to end up in the outer two cells at the end of meiosis (Burt and Trivers, 2009).

In most homosporous ferns, spores are produced from one archesporial cell, which goes through four rounds of mitosis to produce 16 cells; then, these mother cells go through one round of meiosis to produce 64 viable spores (Manton, 1950). Because all meiotic products survive, meiotic drive (as it exists in angiosperms) does not occur in homosporous plants. Therefore, there is no selection pressure on chromosome size or composition during gamete formation, although it may occur at other points such as germline mitosis (Hastings, 1991; Clark et al., 2016). The effect of TRD on genome composition in heterosporous plants could be part of the reduction of chromosome number and genome size in angiosperms. Comparative analyses are needed to measure the extent and affect of TRD on genome composition in homosporous and heterosporous plants.

The presence of TRD in homosporous ferns could be tested using reduced representation sequencing such as restriction site-associated DNA sequencing (RADseq). Two parental species could be crossed to form a hybrid, then spores from the hybrid germinated. Using RADseq, the parents, hybrid, and gametophytes derived from the hybrid would be genotyped. Because ferns have an independent gametophyte stage, meiotic products can be assessed directly without the need for a test cross. Thus, progeny arrays of gametophytes could be genotyped to estimate meiotic product ratios and determine if certain alleles are being preferentially transmitted.

Finally, there are several natural study systems to leverage for work on genomic change following a transition to heterospory. As mentioned previously, there are three independent evolutions of heterospory in land plants: heterosporous ferns (Salviniales), lycophytes (Selaginellales and Isoetales), and the seed plants. The heterosporous ferns and lycophytes both have sister homosporous lineages that would be an ideal comparative system. By first using these two groups to understand how heterospory influence genome evolution in spore-dispersed plants, we could build a foundation to compare homosporous fern genomes to heterosporous seed plants genomes. Another natural system for exploring the evolution of heterospory is *Pteris platyzomopsis* (Pteridaceae). This fern appears to be an example of incipient heterospory; it produces spores in two size classes and has dioecious gametophytes, although egg-producing gametophytes will also produce antheridia after a certain period of time (Tryon, 1964). *Pteris platyzomopsis* has a base chromosome number of $n = 38$ (Tryon and Vida, 1967), which is similar to other homosporous species, but nothing is known about its genome size or structure. More research is also needed on the life history of this species. Incorporating *P. platyzomopsis* in comparative work will be important to understand the evolution of heterospory at the genomic level.

CONCLUSION

The disparity between chromosome number in homosporous and heterosporous plants is one that has challenged scientists for decades (e.g., Klekowski and Baker, 1966). To conduct the necessary comparative analyses, more high-quality genomes from homosporous ferns and lycophytes are needed, which is something many authors have been seeking for almost 20 years (Pryer et al., 2002; Sessa et al., 2014; Wolf et al., 2015; Kuo and Li, 2019). Today, however, our genomic resources are growing rapidly and in a few years, there may be enough homosporous genomes to begin comparative analyses with heterosporous lineages (Kuo and Li, 2019; Szövényi et al., 2021). When investigating broad-scale patterns in vascular plant evolution, it is important to include an evolutionary perspective. In this review, we examined the multiple origins of heterospory in plants and considered which traits might be affecting their chromosome

numbers, genome size, and genome composition. A combination of new data and new ways of looking at the problem may determine which factors are involved and which assumptions we have overlooked.

AUTHOR CONTRIBUTIONS

PW and SK conceived of the project. CR performed the data analyses and figure design. All authors participated in writing and editing the manuscript.

FUNDING

This research was funded by award number 1911459 from the National Science Foundation to PW.

ACKNOWLEDGMENTS

We thank Jacob Suissa, Blaine Marchant, Rijan Dhakal, Chris Haufler, and Tom Ranker for their thoughtful comments on the manuscript.

REFERENCES

Andersson-Kottö, I. (1927). Note on some characters in ferns subject to Mendelian inheritance. *Hereditas* 9, 157–168. doi: 10.1111/j.1601-5223.1927.tb03517.x

Andersson-Kottö, I. (1929). A genetical investigation in *Scolopendrium vulgare*. *Hereditas* 12, 109–177. doi: 10.1111/j.1601-5223.1929.tb02500.x

Andersson-Kottö, I. (1938). "Genetics," in *Manual of Pteridology*, eds F. Verdoorn and A. H. G. Alston (Berlin: Springer), 284–302.

Bainard, J. D., Henry, T. A., Bainard, L. D., and Newmaster, S. G. (2011). DNA content variation in monilophytes and lycophytes: large genomes that are not endopolyploid. *Chromosome Res.* 19, 763–775. doi: 10.1007/s10577-011-9228-1

Baniaga, A. E., and Barker, M. S. (2019). Nuclear genome size is positively correlated with median LTR-RT insertion time in fern and lycophyte genomes. *Am. Fern. J.* 109, 248–266. doi: 10.1640/0002-8444-109.3.248

Banks, J. A., Nishiyama, T., Hasebe, M., Bowman, J. L., Gribskov, M., dePamphilis, C., et al. (2011). The *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants. *Science* 332, 960–963. doi: 10.1126/science.1203810

Barker, M. S. (2013). "Karyotype and genome evolution in pteridophytes," in *Plant Genome Diversity Volume 2: Physical Structure, Behaviour and Evolution of Plant Genomes*, eds J. Greilhuber, J. Dolezel, and J. F. Wendel (Vienna: Springer Vienna), 245–253. doi: 10.1007/978-3-7091-1160-4_15

Barker, M. S., and Wolf, P. G. (2010). Unfurling fern biology in the genomics age. *Bioscience* 60, 177–185. doi: 10.1525/bio.2010.60.3.4

Bateman, R. M., and DiMichele, W. A. (1994). Heterospory: the most iterative key innovation in the evolutionary history of the plant kingdom. *Biol. Rev. Camb. Philos. Soc.* 69, 345–417. doi: 10.1111/j.1469-185X.1994.tb01276.x

Blackburn, K. B., and Harrison, J. W. H. (1924). A preliminary account of the chromosomes and chromosome behaviour in the Salicaceae. *Ann. Bot.* 38, 361–378. doi: 10.1093/oxfordjournals.aob.a089900

Bowers, J. E., and Paterson, A. H. (2021). Chromosome number is key to longevity of polyploid lineages. *New Phytol.* 231, 19–28. doi: 10.1111/nph.17361

Bowman, J. L., Kohchi, T., Yamato, K. T., Jenkins, J., Shu, S., Ishizaki, K., et al. (2017). Insights into land plant evolution garnered from the *Marchantia polymorpha* genome. *Cell* 171, 287.e–304.e. doi: 10.1016/j.cell.2017.09.030

Burt, A., and Trivers, R. (2009). *Genes in Conflict: The Biology of Selfish Genetic Elements*. Cambridge, MA: Harvard University Press.

Carta, A., Bedini, G., and Peruzzi, L. (2020). A deep dive into the ancestral chromosome number and genome size of flowering plants. *New Phytol* 22, 1097–1106. doi: 10.1111/nph.16668

Chapman, R. H., Klekowski, E. J. Jr., and Selander, R. K. (1979). Homoeologous heterozygosity and recombination in the fern *Pteridium aquilinum*. *Science* 204, 1207–1209. doi: 10.1126/science.204.4398.1207

Chester, M., Gallagher, J. P., Symonds, V. V., Cruz da Silva, A. V., Mavrodiev, E. V., Leitch, A. R., et al. (2012). Extensive chromosomal variation in a recently formed natural allopolyploid species, *Tragopogon miscellus (Asteraceae)*. *Proc. Natl. Acad. Sci. U.S.A.* 109, 1176–1181. doi: 10.1073/pnas.1112041109

Clark, J., Hidalgo, O., Pellicer, J., Liu, H., Marquardt, J., Robert, Y., et al. (2016). Genome evolution of ferns: evidence for relative stasis of genome size across the fern phylogeny. *New Phytol.* 210, 1072–1082. doi: 10.1111/nph.13833

Clausen, J. (1927). Chromosome number and the relationship of species in the genus *Viola*. *Ann. Bot.* 41, 677–714. doi: 10.1093/oxfordjournals.aob.a090098

Clausen, R. E., and Goodspeed, T. H. (1925). Interspecific hybridization in *Nic*otiana. II. A tetraploid GLUTINOSA-TABACUM hybrid, an experimental verification of Winge's hypothesis. *Genetics* 10, 278–284. doi: 10.1093/genetics/10.3.278

Cremer, T., and Cremer, C. (1988). Centennial of Wilhelm Waldeyer's introduction of the term "chromosome" in 1888. *Cytogeneics Cell Genet.* 48, 65–67. doi: 10.1159/000132590

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

Flemming, W. (1965). Contributions to the knowledge of the cell and its vital processes. *J. Cell Biol.* 25, 3–69.

Fujiwara, T., Liu, H., Meza-Torres, E. I., Morero, R. E., Vega, A. J., Liang, Z., et al. (2021). Evolution of genome space occupation in ferns: linking genome diversity and species richness. *Ann. Bot.* mcab094. doi: 10.1093/aob/mcab094

Gastony, G. J., and Gottlieb, L. D. (1985). Genetic variation in the homosporous fern *Pellaea andromedifolia*. *Am. J. Bot.* 72, 257–267. doi: 10.1002/j.1537-2197.1985.tb08290.x

Gates, R. R., Sheffield, F. M. L., and Farmer, J. B. (1929). VII. Chromosome linkage in certain *Oenothera* hybrids. *Philos. Trans. R. Soc. Lond. B* 217, 367–394.

Glover, N. M., Redestig, H., and Dessimoz, C. (2016). Homoeologs: what are they and how do we infer them? *Trends Plant Sci.* 21, 609–621. doi: 10.1016/j.tplants.2016.02.005

Hastings, I. M. (1991). Germline selection: population genetic aspects of the sexual/asexual life cycle. *Genetics* 129, 1167–1176. doi: 10.1093/genetics/129.4.1167

Haufler, C. H. (1987). Electrophoresis is modifying our concepts of evolution in homosporous pteridophytes. *Am. J. Bot.* 74, 953–966. doi: 10.1002/j.1537-2197.1987.tb08700.x

Haufler, C. H. (1989). Towards a synthesis of evolutionary modes and mechanisms in homosporous pteridophytes. *Biochem. Syst. Ecol.* 17, 109–115. doi: 10.1016/0305-1978(89)90068-9

Haufler, C. H. (2002). Homospory 2002: an odyssey of progress in pteridophyte genetics and evolutionary biology: ferns and other homosporous vascular plants have highly polyploid chromosome numbers, but they express traits following diploid models and, although capable of extreme inbreeding, are predominantly outcrossing. *Bioscience* 52, 1081–1093. doi: 10.1641/0006-3568(2002)052[1081:haoopi]2.0.co;2

Haufler, C. H., and Soltis, D. E. (1984). Obligate outcrossing in a homosporous fern: field confirmation of a laboratory prediction. *Am. J. Bot.* 71, 878–881. doi: 10.1002/j.1537-2197.1984.tb14153.x

Haufler, C. H., and Soltis, D. E. (1986). Genetic evidence suggests that homosporous ferns with high chromosome numbers are diploid. *Proc. Natl. Acad. Sci. U.S.A.* 83, 4389–4393. doi: 10.1073/pnas.83.12.4389

Haufler, C. H., Pryer, K. M., Schuettpelz, E., Sessa, E. B., Farrar, D. R., Moran, R., et al. (2016). Sex and the single gametophyte: revising the homosporous vascular plant life cycle in light of contemporary research. *Bioscience* 66, 928–937. doi: 10.1093/biosci/biw108

Hauk, W. D., and Haufler, C. H. (1999). Isozyme variability among cryptic species of *Botrychium* subgenus *Botrychium* (Ophioglossaceae). *Am. J. Bot.* 86, 614–633. doi: 10.2307/2656570

Hickok, L. G. (1979). A cytological study of intraspecific variation in *Ceratopteris thalictroides*. *Can. J. Bot.* 57, 1694–1700. doi: 10.1139/b79-207

Hickok, L. G., and Klekowski, E. J. Jr. (1973). Abnormal reductional and non-reductional meiosis in *Ceratopteris*: alternatives to homozygosity and hybrid sterility in homosporous ferns. *Am. J. Bot.* 60, 1010–1022. doi: 10.1002/j.1537-2197.1973.tb06002.x

Huang, L. O., Labbe, A., and Infante-Rivard, C. (2013). Transmission ratio distortion: review of concept and implications for genetic association studies. *Hum. Genet.* 132, 245–263. doi: 10.1007/s00439-012-1257-0

Klekowski, E. J. (1969). Reproductive biology of the *Pteridophyta*. II. Theoretical considerations. *Bot. J. Linn. Soc.* 62, 347–359. doi: 10.1111/j.1095-8339.1969.tb01972.x

Klekowski, E. J. (1972). Genetical features of ferns as contrasted to seed plants. *Ann. Mol. Bot. Gard.* 59, 138–151. doi: 10.2307/2394749

Klekowski, E. J. (1973). Sexual and subsexual systems in homosporous pteridophytes: a new hypothesis. *Am. J. Bot.* 60, 535–544. doi: 10.1002/j.1537-2197.1973.tb05955.x

Klekowski, E. J. (1976). Homoeologous chromosome pairing in ferns. *Curr. Chrom. Res* 82, 175–184.

Klekowski, E. J. Jr., and Hickok, L. G. (1974). Nonhomologous chromosome pairing in the fern *Ceratopteris*. *Am. J. Bot.* 61, 422–432. doi: 10.1002/j.1537-2197.1974.tb12261.x

Klekowski, E. J., and Baker, H. G. (1966). Evolutionary significance of polyploidy in the *pteridophyta*. *Science* 153, 305–307. doi: 10.1126/science.153.3733.305

Kuo, L.-Y., and Li, F.-W. (2019). A roadmap for fern genome sequencing. *Am. Fern J.* 109, 212–223. doi: 10.1640/0002-8444-109.3.212

Lang, W. H. (1923). On the genetic analysis of a heterozygotic plant of *Scolopendrium vulgare*. *J. Genet.* 13, 167–175. doi: 10.1007/BF02983052

Leitch, A. R., and Leitch, I. J. (2012). Ecological and genetic factors linked to contrasting genome dynamics in seed plants. *New Phytol.* 194, 629–646. doi: 10.1111/j.1469-8137.2012.04105.x

Leitch, I. J., and Bennett, M. D. (2004). Genome downsizing in polyploid plants. *Biol. J. Linn. Soc. Lond.* 82, 651–663. doi: 10.1111/j.1095-8312.2004.00349.x

Leitch, I. J., and Leitch, A. R. (2013). "Genome size diversity and evolution in land plants," in *Plant Genome Diversity Volume 2: Physical Structure, Behaviour and Evolution of Plant Genomes*, eds J. Greilhuber, J. Dolezel, and J. F. Wendel (Vienna: Springer Vienna), 307–322. doi: 10.1007/978-3-7091-1160-4_19

Li, F.-W., Brouwer, P., Carretero-Paulet, L., Cheng, S., de Vries, J., Delaux, P.-M., et al. (2018). Fern genomes elucidate land plant evolution and cyanobacterial symbioses. *Nat. Plants* 4, 460–472. doi: 10.1038/s41477-018-0188-8

Li, F.-W., Nishiyama, T., Waller, M., Frangedakis, E., Keller, J., Li, Z., et al. (2020). *Anthoceros* genomes illuminate the origin of land plants and the unique biology of hornworts. *Nat. Plants* 6, 259–272. doi: 10.1038/s41477-020-0618-2

Li, Z., McKibben, M. T. W., Finch, G. S., Blischak, P. D., Sutherland, B. L., and Barker, M. S. (2020). Patterns and processes of diploidization in land plants. *Annu. Rev. Plant Biol.* 72, 387–410. doi: 10.5281/zenodo.3964504

Liang, Z., Geng, Y., Ji, C., Du, H., Wong, C. E., Zhang, Q., et al. (2020). *Mesostigma viride* genome and transcriptome provide insights into the origin and evolution of Streptophyta. *Adv. Sci.* 7:1901850. doi: 10.1002/advs.201901850

Lim, K. Y., Soltis, D. E., Soltis, P. S., Tate, J., Matyasek, R., Srubarova, H., et al. (2008). Rapid chromosome evolution in recently formed polyploids in *Tragopogon (Asteraceae)*. *PLoS One* 3:e3353. doi: 10.1371/journal.pone.0003353

Liu, H.-M., Ekrt, L., Koutecky, P., Pellicer, J., Hidalgo, O., Marquardt, J., et al. (2019). Polyploidy does not control all lineage-specific average chromosome length constrains genome size evolution in ferns. *J. Syt. Evol.* 57:12525. doi: 10.1111/jse.12525

Lutz, A. M. (1907). A preliminary note on the chromosomes of *Oenothera lamarckiana* and one of its mutants. O. Gigas. *Science* 26, 151–152. doi: 10.1126/science.26.657.151

Manton, I. (1950). *Problems of Cytology In The Pteridophyta*. Cambridge, MA: Cambridge University Press.

Manton, I., and Sledge, W. A. (1954). Observations on the cytology and taxonomy of the pteridophyte flora of Ceylon. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 238, 127–185. doi: 10.1098/rstb.1954.0008

Marchant, D. B. (2019). Ferns with benefits: incorporating *Ceratopteris* into the genomics era. *Am. Fern. J.* 193, 183–191. doi: 10.1640/0002-8444-109.3.183

Marchant, D. B., Sessa, E. B., Wolf, P. G., Heo, K., Barbazuk, W. B., Soltis, P. S., et al. (2019). The C-Fern (*Ceratopteris richardii*) genome: insights into plant genome evolution with the first partial homosporous fern genome assembly. *Sci. Rep.* 9:18181. doi: 10.1038/s41598-019-53968-8

Nakazato, T., Barker, M. S., Rieseberg, L. H., and Gastony, G. J. (2008). "Evolution of the nuclear genome of ferns and lycophytes," in *Biology and evolution of ferns and lycophytes*, eds T. A. Ranker and C. H. Haufler (Cambridge, MA: Cambridge University Press), 175–198. doi: 10.1017/CBO9780511541827.008

Nakazato, T., Jung, M.-K., Housworth, E. A., Rieseberg, L. H., and Gastony, G. J. (2006). Genetic map-based analysis of genome structure in the homosporous fern *Ceratopteris richardii*. *Genetics* 173, 1585–1597. doi: 10.1534/genetics.106.055624

Noyes, R. D., and Givens, A. D. (2013). Auantitative assessment of megasporogenesis for the facultative apomicts *Erigeron annuus* and *Erigeron strigosus* (*Asteraceae*). *Int. J. Plant Sci.* 174, 1239–1250. doi: 10.1086/673243

Nystedt, B., Street, N. R., Wetterbom, A., Zuccolo, A., Lin, Y.-C., Scofield, D. G., et al. (2013). The Norway spruce genome sequence and conifer genome evolution. *Nature* 497, 579–584. doi: 10.1038/nature12211

Ohno, S. (1984). Conservation of linkage relationships between genes as the underlying theme of karyological evolution in mammals. *Chromosome Evol. Eukaryotic Groups* 2, 1–11.

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

Pardo-Manuel, de Villena, F., and Sapienza, C. (2001). Nonrandom segregation during meiosis: the unfairness of females. *Mamm. Genome* 12, 331–339. doi: 10.1007/s003350040003

Pellicer, J., and Leitch, I. J. (2020). The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. *New Phytol.* 226, 301–305. doi: 10.1111/nph.16261

PPG, I. (2016). A community-derived classification for extant lycophytes and ferns: PPG I. *J. Sytemat. Evol.* 54, 563–603. doi: 10.1111/jse.12229

Pryer, K. M., Schneider, H., Zimmer, E. A., and Ann Banks, J. (2002). Deciding among green plants for whole genome studies. *Trends Plant Sci.* 7, 550–554. doi: 10.1016/s1360-1385(02)02375-0

Pryer, K. M., Schuettpelz, E., Wolf, P. G., Schneider, H., Smith, A. R., and Cranfill, R. (2004). Phylogeny and evolution of ferns (*Monilophytes*) with a focus on the early leptosporangiate divergences. *Am. J. Bot.* 91, 1582–1598. doi: 10.3732/ajb.91.10.1582

Pryer, K. M., Smith, A. R., and Skog, J. E. (1995). Phylogenetic relationships of extant ferns based on evidence from morphology and rbcl sequences. *Am. Fern J.* 85, 205–282. doi: 10.1186/s12862-015-0400-7

Ramsey, J., and Schemske, D. W. (2002). Neopolyploidy in flowering plants. *Annu. Rev. Ecol. Syst.* 33, 589–639. doi: 10.1146/annurev.ecolsys.33.010802.150437

Rensing, S. A., Lang, D., Zimmer, A. D., Terry, A., Salamov, A., Shapiro, H., et al. (2008). The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science* 319, 64–69. doi: 10.1126/science.1150646

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

Schmerler, S., and Wessel, G. M. (2011). Polar bodies-More a lack of understanding than a lack of respect. *Mol. Repro. Dev.* 78, 3–8. doi: 10.1002/mrd.21266

Schubert, I., and Lysak, M. A. (2011). Interpretation of karyotype evolution should consider chromosome structural constraints. *Trends Genet.* 27, 207–216. doi: 10.1016/j.tig.2011.03.004

Schubert, I., and Oud, J. L. (1997). There is an upper limit of chromosome size for normal development of an organism. *Cell* 88, 515–520. doi: 10.1016/S0092-8674(00)81891-7

Sessa, E. B., Banks, J. A., Barker, M. S., Der, J. P., Duffy, A. M., Graham, S. W., et al. (2014). Between two fern genomes. *Gigascience* 3:15. doi: 10.1186/2047-217X-3-15

Shakirov, E. V., and Shippen, D. E. (2012). *Selaginella moellendorffii* telomeres: conserved and unique features in an ancient land plant lineage. *Front. Plant Sci.* 3:161. doi: 10.3389/fpls.2012.00161

Šimoníková, D., Nìmeèková, A., Karafiátová, M., Uwimana, B., Swennen, R., Doležel, J., et al. (2019). Chromosome painting facilitates anchoring reference genome sequence to chromosomes in situ and integrated karyotyping in banana (*Musa* spp.). *Front. Plant Sci.* 10:1503. doi: 10.3389/fpls.2019.01503

Skovsted, A. (1935). Cytological studies in cotton. *J. Genet.* 30, 397–405. doi: 10.1007/bf02982248

Soltis, D. E., and Soltis, P. S. (1986). Electrophoretic evidence for inbreeding in the fern *Botrychium virginianum* (Ophioglossaceae). *Am. J. Bot.* 73, 588–592. doi: 10.1002/j.1537-2197.1986.tb12078.x

Strasburger, E. (1894). The periodic reduction of the number of the chromosomes in the life-history of living organisms. *Ann. Bot.* 8, 281–316. doi: 10.1093/oxfordjournals.aob.a090708

Strasburger, E. (1910). Chromosomenzahl. *Flora BD* 100, 398–446.

Szövényi, P., Gunadi, A., and Li, F.-W. (2021). Charting the genomic landscape of seed-free plants. *Nat. Plants* 7, 554–565. doi: 10.1038/s41477-021-00888-z

Takuno, S., Ran, J.-H., and Gaut, B. S. (2016). Evolutionary patterns of genic DNA methylation vary across land plants. *Nat. Plants* 2:15222. doi: 10.1038/nplants.2015.222

The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815. doi: 10.1038/35048692

Tryon, A. F. (1964). Platyzoma - a Queensland fern with incipient heterospory. *Am. J. Bot.* 51, 939–942. doi: 10.1002/j.1537-2197.1964.tb06721.x

Tryon, A. F., and Vida, G. (1967). *Platyzoma*: a new look at an old link in ferns. *Science* 156, 1109–1110. doi: 10.1126/science.156.3778.1109

Van de Peer, Y., Maere, S., and Meyer, A. (2009). The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.* 10, 725–732. doi: 10.1038/nrg2600

Van de Peer, Y., Mizrachi, E., and Marchal, K. (2017). The evolutionary significance of polyploidy. *Nat. Rev. Genet.* 18, 411–424. doi: 10.1038/nrg.2017.26

Volkmann, D., Baluška, F., and Menzel, D. (2012). Eduard Strasburger (1844-1912): founder of modern plant cell biology. *Protoplasma* 249, 1163–1172. doi: 10.1007/s00709-012-0406-6

Wagner, W. H. Jr., and Wagner, F. S. (1979). Polyploidy in pteridophytes. *Basic Life Sci.* 13, 199–214. doi: 10.1007/978-1-4613-3069-1_11

Weiss-Schneeweiss, H., and Schneeweiss, G. M. (2013). "Karyotype diversity and evolutionary trends in angiosperms," in *Plant Genome Diversity Volume 2: Physical Structure, Behaviour and Evolution of Plant Genomes*, eds J. Greilhuber, J. Dolezel, and J. F. Wendel (Vienna: Springer Vienna), 209–230. doi: 10.1007/978-3-7091-1160-4_13

Wendel, J. F., Jackson, S. A., Meyers, B. C., and Wing, R. A. (2016). Evolution of plant genome architecture. *Genome Biol.* 17:37. doi: 10.1186/s13059-016-0908-1

Winge, O. (1917). The chromosomes: their numbers and general importance. *Comptes Rend. Travaux Lab. Carlsberg* 13, 131–175.

Winkelmann, A. (2007). Wilhelm von Waldeyer-Hartz (1836-1921): an anatomist who left his mark. *Clin. Anat.* 20, 231–234. doi: 10.1002/ca.20400

Wolf, P. G., Haufler, C. H., and Sheffield, E. (1987). Electrophoretic evidence for genetic diploidy in the bracken fern (*Pteridium aquilinum*). *Science* 236, 947–949. doi: 10.1126/science.236.4804.947

Wolf, P. G., Sessa, E. B., Marchant, D. B., Li, F.-W., Rothfels, C. J., Sigel, E. M., et al. (2015). An exploration into fern genome space. *Genome Biol. Evol.* 7, 2533–2544. doi: 10.1093/gbe/evv163

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

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

Copyright © 2022 Kinosian, Rowe and Wolf. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Plant Science

PERSPECTIVE
published: 14 March 2022
doi: 10.3389/fpls.2022.863076

Loss of Plastid Developmental Genes Coincides With a Reversion to Monoplastidy in Hornworts

Alexander I. MacLeod[1], Parth K. Raval[1], Simon Stockhorst[1], Michael R. Knopp[1], Eftychios Frangedakis[2] and Sven B. Gould[1]**

[1] Institute for Molecular Evolution, Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Germany, [2] Department of Plant Sciences, University of Cambridge, Cambridge, United Kingdom

The first plastid evolved from an endosymbiotic cyanobacterium in the common ancestor of the Archaeplastida. The transformative steps from cyanobacterium to organelle included the transfer of control over developmental processes, a necessity for the host to orchestrate, for example, the fission of the organelle. The plastids of almost all embryophytes divide independently from nuclear division, leading to cells housing multiple plastids. Hornworts, however, are monoplastidic (or near-monoplastidic), and their photosynthetic organelles are a curious exception among embryophytes for reasons such as the occasional presence of pyrenoids. In this study, we screened genomic and transcriptomic data of eleven hornworts for components of plastid developmental pathways. We found intriguing differences among hornworts and specifically highlight that pathway components involved in regulating plastid development and biogenesis were differentially lost in this group of bryophytes. Our results also confirmed that hornworts underwent significant instances of gene loss, underpinning that the gene content of this group is significantly lower than other bryophytes and tracheophytes. In combination with ancestral state reconstruction, our data suggest that hornworts have reverted back to a monoplastidic phenotype due to the combined loss of two plastid division-associated genes, namely, ARC3 and FtsZ2.

Keywords: plastid evolution, bryophytes, hornworts, plant terrestrialization, plastid division

OPEN ACCESS

Edited by:
Michael Eric Schranz, Wageningen University & Research, Netherlands

Reviewed by:
Yoshihisa Hirakawa, University of Tsukuba, Japan

****Correspondence:***
Alexander I. MacLeod
macleod@hhu.de
Sven B. Gould
gould@hhu.de

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

Received: *26 January 2022*
Accepted: *04 February 2022*
Published: *14 March 2022*

Citation:
MacLeod AI, Raval PK, Stockhorst S, Knopp MR, Frangedakis E and Gould SB (2022) Loss of Plastid Developmental Genes Coincides With a Reversion to Monoplastidy in Hornworts. Front. Plant Sci. 13:863076. doi: 10.3389/fpls.2022.863076

INTRODUCTION

Hornworts are a unique group of bryophytes, the monophyletic non-vascular sister lineage to all vascular land plants (Harris et al., 2020). The phylogenetic position of hornworts and their putative phenotypic resemblance to what one might consider to represent the last common ancestor of all land plants make them an attractive model for evo-devo studies linked to events such as plant terrestrialization (Frangedakis et al., 2020). Hornworts are the only group of land plants known to form a pyrenoid, a unique carbon-concentrating mechanism (CCM), otherwise common in algae; however, these CCMs are not present in all hornworts and are hence a poor taxonomic marker (Villarreal and Renner, 2012; **Supplementary Figure 1**).

Hornworts are one of the only groups of embryophytes that have not escaped the monoplastidic bottleneck. This is a phenomenon associated with plastid origin and the organelle's integration into the host cell cycle, which constrains the majority of algae from possessing multiple plastids per cell (de Vries and Gould, 2018). One consequence is that the only plastids—of which there are five types in embryophytes (Jarvis and López-Juez, 2013)—hornwort cells house are chloroplasts,

whose size and morphology vary across genera (Vaughn et al., 1992; Raven and Edwards, 2014; Li et al., 2017). To address the reason, we screened the genomes and annotated transcriptomes of ten hornwort species to identify the presence/absence of genes that play key roles in regulating plastid development, such as those involved in protein import into the chloroplast, thylakoid biogenesis, and chloroplast division (Jarvis and López-Juez, 2013). We highlight key differences between the developmental plastid biology of hornworts and other established model organisms in the terrestrial clade. Furthermore, we argue that the major changes in plastid biology, that not only coincided with major checkpoints in the evolutionary history of hornworts but also facilitated them, are a consequence of multiple instances of gene loss observed in this unique group of embryophytes.

HORNWORTS UNDERWENT SIGNIFICANT INSTANCES OF GENE LOSS

We used the BUSCO version 5.2.2 (Manni et al., 2021) software to estimate the gene content of hornworts and compared them with other bryophyte and tracheophyte (vascular plant) outgroups (**Supplementary Table 1**). We found that the gene content of hornworts is significantly lower than tracheophytes and other bryophytes (ANOVA; $F = 129.5$; d.f. = 2,30; $p < 0.001$), thereby suggesting that hornwort diversification and speciation were accompanied by significant instances of gene loss (**Supplementary Figure 2**), even more than what is observed for bryophytes in general (Harris et al., 2021).

FULL CONSERVATION OF TRANSLOCON OF THE OUTER ENVELOPE OF THE CHLOROPLAST BUT ONLY PARTIAL CONSERVATION OF TRANSLOCON OF THE INNER ENVELOPE OF THE CHLOROPLAST IN HORNWORT CHLOROPLASTS

The vast majority of plastid proteins are encoded by the nuclear genome, and after their synthesis in the cytosol, are imported into the plastid by the translocon of the outer/inner envelope of the chloroplast (TOC/TIC) complex (Richardson and Schnell, 2020). Embryophytes have evolved the most sophisticated TOC/TIC complexes (Gould et al., 2008; Knopp et al., 2020) and our data confirm that the hornwort TOC complex is comprised of the same key proteins that are found in other embryophytes, mainly TOC75, TOC34, and TOC159 (Richardson and Schnell, 2020; **Figure 1**). The recycling of major TOC components is regulated by the RING-type ubiquitin E3 ligase SP1, which targets these proteins for proteasomal degradation (Ling et al., 2012; **Figure 1B**).

The TIC complex of embryophytes is comprised of a 1 MDa multimer that forms a pore that receives precursor proteins from the TOC complex in the intermembrane space (IMS) and finally mediates their passage to the stroma (Nakai, 2015a; Richardson and Schnell, 2020; **Figure 1**). The presence/absence of TOC/TIC components reveals no pattern with regard to mono-/polyplastidy or presence/absence of a pyrenoid (**Figure 1A** and **Supplementary Figure 1**). However, some TIC components appear to have undergone differential loss in some hornwort taxa (**Figure 1A**), most notably TIC21, TIC22, YCF1 (TIC214), and maybe even TIC20 in *Leiosporoceros dussii*. The latter species is the only member of our surveyed taxa that lacks a TIC20 ortholog (**Figure 1A**).

YCF1/TIC214, the only TOC/TIC component encoded by the plastid genome and unique to the green lineage, is absent in a significant number of hornworts (**Figure 1A**), such as in *Nothoceros aenigmaticus*, for which also the plastid genome is available (Villarreal et al., 2013).

DIFFERENTIAL LOSS OF AN ANCIENT THYLAKOID DEVELOPMENTAL PATHWAY IN MOST HORNWORTS

Thylakoid proteomes contain the bulk of photosynthesis-related proteins of plant cells (Xu et al., 2021). After their import *via* TOC/TIC, thylakoid proteins are recognized and sorted *via* one of three main pathways, the components of which are predominantly derived from the cyanobacterial endosymbiont or inserted spontaneously (Xu et al., 2021; **Figure 1**).

The chloroplast secretory (cpSec) pathway is involved in importing unfolded proteins to the thylakoid lumen. Powered by the motor protein cpSecA, unfolded subunits pass through a pore formed by cpSecY and cpSecE (Xu et al., 2021; **Figure 1B**). While the presence/absence of cpSec components reveals no pattern with regard to mono-/polyplastidy or presence/absence of a pyrenoid, half of surveyed hornworts lack cpSecE orthologs, with this distribution not showing any unique phylogenetic pattern (**Figure 1A** and **Supplementary Figure 1**).

The chloroplast twin-arginine translocation (cpTat) pathway can import folded proteins and is powered by the thylakoid's proton motive force (PMF; Xu et al., 2021). In those hornworts, for which we identified the cpTat pathway, it is comprised of three proteins, namely, Tha4, TatC, and Hcf106 (**Figure 1**). Precursor proteins initially bind to a TatC-Hcf106 complex. Tha4 is subsequently recruited *via* the action of the PMF, undergoing a conformational change, leading to the passage of the precursor protein (Xu et al., 2021; **Figure 1B**). The presence/absence of cpTat components reveals no pattern with regard to mono-/polyplastidy or presence/absence of a pyrenoid; however, the cpTat pathway seems only to be encoded by the Anthocerotaceae, having been differentially lost in other hornwort families (**Figure 1A**).

The third main pathway involved in sorting proteins for thylakoid biogenesis is the chloroplast signal recognition particle (cpSRP) pathway. This translocation complex is involved in targeting specifically light harvesting complex proteins (LHCPs) to the thylakoid membrane (Xu et al., 2021; **Figure 1B**). LHCP integration is initiated when a rudimentary LHCP is transferred from the TIC translocon to the SRP43/SRP54 complex by the

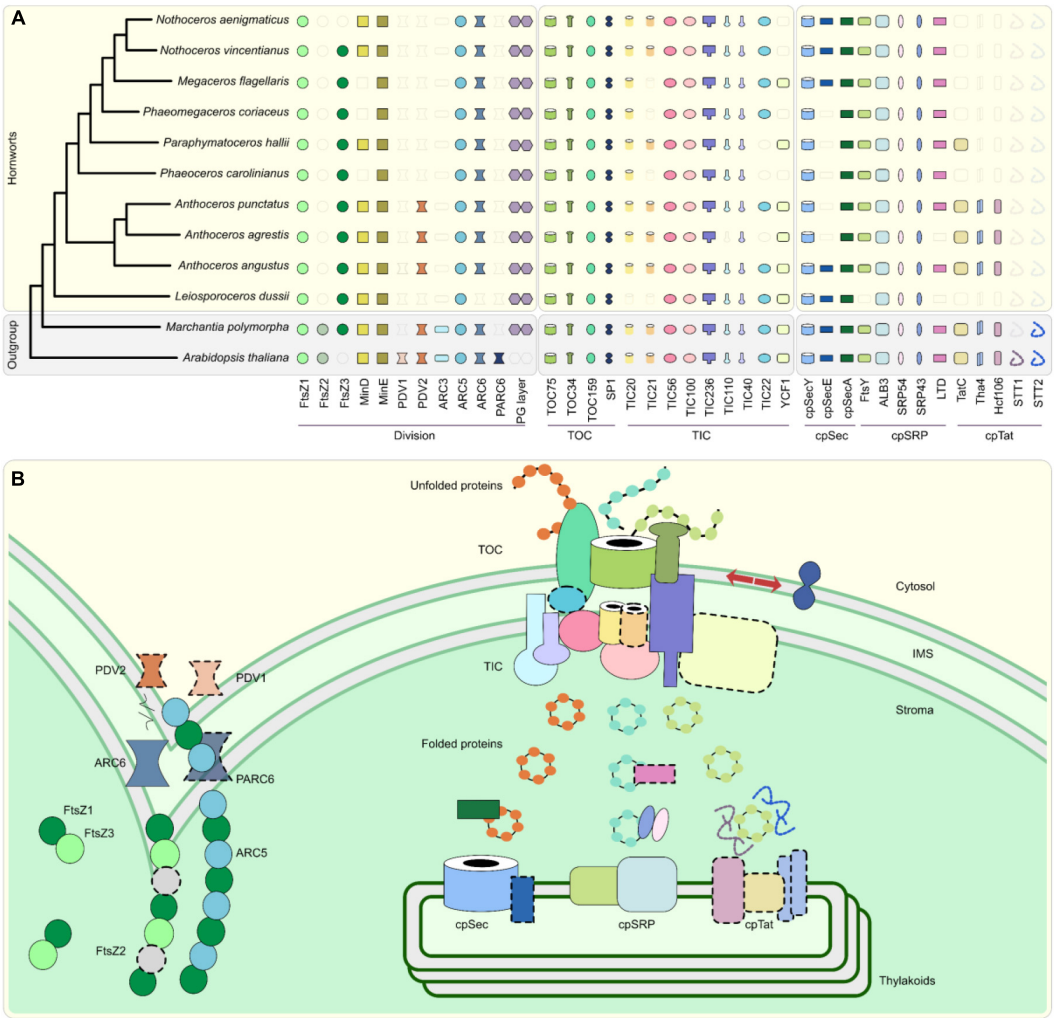

FIGURE 1 | Plastid development and biogenesis in hornworts. **(A)** A presence/absence pattern (PAP) of various plastid developmental components that are sorted into three categories based on whether they are associated with plastid division (PD) and protein translocation across the plastid envelope *via* TOC/TIC or the thylakoid membrane. Transparent icons indicate that no gene could be identified. **(B)** A combined schematic representation of plastid development in embryophytes. Components that are absent from more than two hornworts in our surveyed taxa, or absent in this group altogether, are highlighted by dotted outlines. ARC, accumulation and regulation of chloroplasts; FtsZ, filamentous temperature Z; IMS, intermembrane space; Sec, secretory; SRP, signal recognition particle; Tat; twin arginine translocation; TOC/TIC, translocator of the outer/inner chloroplast membrane; PDV, plastid division. While ARC5 is absent from the *Anthoceros agrestis* Bonn ecotype, which we included in our OrthoFinder analyses as the representative for this species, our reciprocal best hit pipeline confirmed that it is present in the Oxford ecotype, with its gene ID being AagrOXF_evm.TU.utg000081l.174. A maximum likelihood (ML) tree was constructed *via* the IQ-TREE version 2.0.3 software (Minh et al., 2020), using an automated selection model, by concatenating single-copy chloroplast and mitochondrial markers from 65 different hornwort species, and three outgroups (Villarreal and Renner, 2012). Said sequences were aligned with MUSCLE in AliView (Edgar, 2004; Laarson, 2014). Gene trees for orthologs listed on the PAP were generated using the PhyML version 3.0 and IQ-TREE version 2.0.3 softwares using automated selection models (Guindon et al., 2010; Lefort et al., 2017). We used the SHOOT framework (Emms and Kelly, 2021) to extract orthologous sequences from across the Archaeplastida for said trees. We analyzed the genomes and transcriptomes of ten hornworts, along with the genomes of *Arabidopsis thaliana* and *Marchantia polymorpha*, to determine the presence of various components involved in plastid development (Lamesch et al., 2012; Bowman et al., 2017; Leebens-Mack et al., 2019; Li et al., 2020; Zhang et al., 2020; **Supplementary Table 2**). These orthology clusters (orthogroups) were identified using the OrthoFinder version 2.5.4 software (Emms and Kelly, 2015, 2019; **Supplementary Table 2**). To validate orthogroup presence/absence, we checked for reciprocal best hits using DIAMOND (Buchfink et al., 2015). Due to the difficulty in identifying orthologs for the import protein YCF1 in the Archaeplastida (de Vries et al., 2015), we employed a different strategy to identify orthologs for this gene. We extracted established YCF1 sequences from GenBank and UniProt and used them as queries for DIAMOND.

LTD protein. Subsequently, this SRP43/SRP54 complex binds to the FtsY receptor. GTP hydrolysis results in LHCP integration *via* the action of the ALB3 integral translocase (Xu et al., 2021; **Figure 1B**). Our results suggest that the cpSRP pathway is ubiquitous in all hornworts, as the core components of this pathway are present in the vast majority of our surveyed taxa; therefore, presence/absence of cpSRP components reveals no pattern with regard to mono-/polyplastidy or presence/absence of a pyrenoid. However, FtsY is absent in *L. dussii*, and LTD is absent in both *Anthoceros angustus* and *L. dussii*.

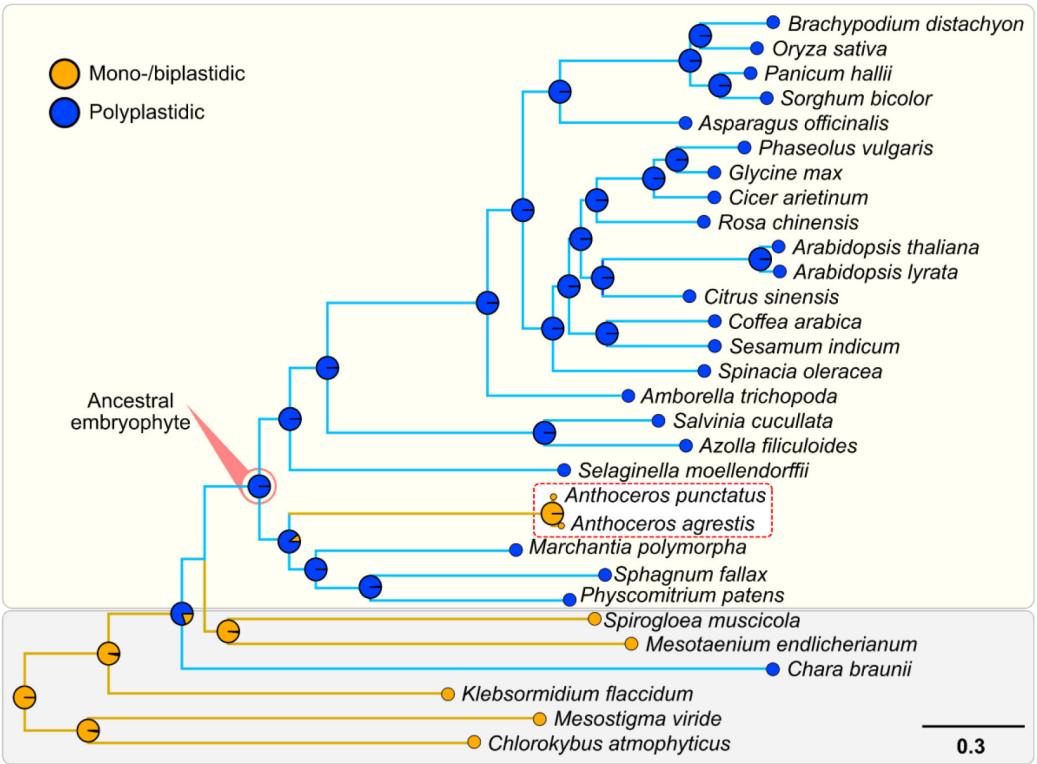

FIGURE 2 | Support for the polyplastidic nature of the ancestral embryophyte and monoplastidic nature of the ancestral hornwort. Pie charts at the nodes display estimates of the probabilities for the plastidic phenotype of the respective most recent common ancestors (MRCAs). Hornworts are highlighted with a white box and a red dotted line. A robust ML species phylogeny of the green lineage was constructed *via* the IQ-TREE version 2.0.3 (Minh et al., 2020), using an automated selection model, by concatenating several housekeeping genes identified with DIAMOND in 34 different streptophytes, seven chlorophytes, and one glaucophyte (**Supplementary Tables 3**, **4**; Buchfink et al., 2015). We used a reciprocal best hit pipeline with DIAMOND (Buchfink et al., 2015), to analyze the genomes of 34 different streptophytes to determine the presence and absence of orthologs involved in plastid division, to estimate the presence/absence of ARC3 and FtsZ2 at various nodes on our tree (**Supplementary Table 5**). Subsequent ASRs were undertaken using the ape function from the Phytools package (Revell, 2012).

LOSS OF PLASTID DIVISION COMPONENTS COINCIDES WITH MONOPLASTIDY IN HORNWORTS

Plastid division in bryophytes is achieved by three components, namely, the outer and inner rings and most likely the peptidoglycan (PG) layer (**Figure 1**). The inner division ring (Z-ring) is comprised of FtsZ1, FtsZ2, and FtsZ3, while the outer division ring comprises ARC5 and FtsZ3 (Osteryoung and Pyke, 2014; Grosche and Rensing, 2017; **Figure 1B**). Z-ring and outer ring synchronization are achieved *via* an interplay of ARC6 and PDV2 (Osteryoung and Pyke, 2014). The PG layer is a relic of the chloroplast's cyanobacterial past, and it might be relevant in regulating chloroplast division in bryophytes and streptophyte algae (Hirano et al., 2016; Grosche and Rensing, 2017).

Hornworts appear to have differentially lost *both* ARC3 and FtsZ2 (**Figure 1A**). This differential loss correlates with this group of bryophytes reverting back to a monoplastidic, or near-monoplastidic, phenotype (**Figure 2** and **Supplementary Figures 3**, **4**; Villarreal and Renner, 2012; Raven and Edwards, 2014; Li et al., 2017). Indeed, previous studies have shown that generating individual gene mutant lines of *ARC3* and *FtsZ2* in *A. thaliana* and the moss *Physcomitrium patens* causes fewer plastids (in the case of *arc3* mutants) or one giant plastid per cell (in the case of *ftsz2* mutants) (Pyke and Leech, 1992; Martin et al., 2009). ARC3 is part of the FtsZ family and unites an FtsZ domain with a C-terminal MORN domain (Zhang et al., 2013).

DISCUSSION

It is evident that hornwort—and bryophyte—emergence and diversification were accompanied by major instances of gene loss (Harris et al., 2021). Our results reinforce this hypothesis, specifically highlighting that the combined loss of certain genes may be responsible for the unique plastid phenotype observed in this group.

The absence of TIC20 in *L. dussii* could be the result of a transcriptome annotation and coverage issues (Cheon et al., 2020), since TIC20 is hypothesized to be a universal protein across the green lineage (Kalanon and McFadden, 2008; de Vries et al., 2015). Should this not be the case, then, maybe YCF1/TIC214 and TIC100 can compensate for TIC20's absence in a unique manner. Some putative absences of YCF1/TIC214 could also be the result of assembly

and/or annotation errors; however, the gene was lost without question in grasses, too (de Vries et al., 2015; Nakai, 2015b). The loss of this import protein does not lead to the loss of the entire import capacity (Bölter and Soll, 2017) and raises the question whether there is a functional, causative correlation between the loss of YCF1/TIC214 across these diverse embryophyte groups.

Considering cpSecE only plays an accessory role in protein translocation by tiling and rotating cpSecY's N-terminal half, its absence in some hornworts indicates that it might not be detrimental to the function of the cpSec pathway (**Figure 1B**; Park et al., 2014). If the cpTat pathway is indeed absent in most hornwort families, then this raises the question on how the thylakoids import folded proteins. Furthermore, all hornworts appear to lack STT proteins (**Figure 1A**), which mediate liquid-liquid phase transitions (LLPTs), allowing for more efficient sorting of cpTat substrates (**Figure 1**; Ouyang et al., 2020). cpTat-related LLPTs hence appear absent in hornworts or are regulated otherwise. The differential loss of FtsY and LTD in *L. dussii* could be a consequence of this species potentially losing TIC20, with this core TIC component being a key LTD interaction partner (Ouyang et al., 2011).

We found that the chloroplasts of all surveyed hornworts possess all the enzymes necessary for PG layer biosynthesis (**Figure 1A**), hinting toward a conserved function similar to that in the moss *P. patens* (Hirano et al., 2016). While ARC3 orthologs are absent in some polyplastidic seedless plants (such as *P. patens* and the lycophyte *Selaginella moellendorffii*), these species then possess orthologs for FtsZ2, which might compensate its loss to some degree (Rensing et al., 2008; Albert et al., 2011; Zhang et al., 2013). This is further supported by an ancestral state reconstruction analysis that demonstrates that the ancestral embryophyte possessed both ARC3 and FtsZ2 and was polyplastidic, the opposite of which is true for the ancestral hornwort (**Figure 2** and **Supplementary Figures 3**, **4**). We predict that the loss of both genes contributed to the monoplastidic nature of hornworts and that reintroducing them might induce a polyplastidic phenotype.

CONCLUSION AND OUTLOOK

We suggest that a consequence of some of plastid-related gene losses, including the combined loss of FtsZ2 and ARC3, resulted in hornworts reverting back to a monoplastidic phenotype, which the embryophyte ancestor was able to escape. If the knockout of ARC3 and FtsZ2 in *A. thaliana* and *P. patens* results in monoplastidic phenotypes, could one reverse evolution by expressing ARC3 and/or FtsZ2 in a hornwort? We anticipate our study to be a starting point for further experiments aimed at deconstructing bryophyte plastid biology and reconstructing new evolutionary hypotheses for outstanding questions in this topic. Next to exploring the monoplastidic bottleneck, hornworts might be able to shed new light on the import of folded proteins into the thylakoid of non-Anthocerotaceae hornworts, or the consequences of a potential TIC20 loss in *L. dussii* and the detailed function of YCF1; which, like all grasses, some hornworts appear to have lost.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

AM undertook the phylogenomic analyses, with help from SS and MK. AM, SG, PR, and EF wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

We are grateful for the support by the DFG (SFB 1208–267205415 and SPP2237–440043394) and VolkswagenStiftung (Life). AM is furthermore supported by the Moore and Simons Initiative grant (9743) of William F. Martin.

ACKNOWLEDGMENTS

We thank William F. Martin for his ongoing support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.863076/full#supplementary-material

REFERENCES

Albert, V. A., Aono, N., and Aoyama, T. (2011). Content Associated With the Evolution of Vascular Plants. *Science* 332, 960–963. doi: 10.1126/science.1203810

Bölter, B., and Soll, J. (2017). Ycf1/Tic214 Is Not Essential for the Accumulation of Plastid Proteins. *Mol. Plant* 10, 219–221. doi: 10.1016/j.molp.2016.10.012

Bowman, J. L., Kohchi, T., Yamato, K. T., Jenkins, J., Shu, S., Ishizaki, K., et al. (2017). Insights into Land Plant Evolution Garnered from the Marchantia polymorpha Genome. *Cell* 171, 287–304.e15. doi: 10.1016/j.cell.2017.09.030

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi: 10.1038/nmeth.3176

Cheon, S., Zhang, J., and Park, C. (2020). Is Phylotranscriptomics as Reliable as Phylogenomics? *Mol. Biol. Evol.* 37, 3672–3683. doi: 10.1093/molbev/msaa181

de Vries, J., and Gould, S. B. (2018). The monoplastidic bottleneck in algae and plant evolution. *J. Cell Sci.* 131:jcs203414. doi: 10.1242/jcs.203414

de Vries, J., Sousa, F. L., Bölter, B., Soll, J., and Goulda, S. B. (2015). YCF1: a green TIC? *Plant Cell* 27, 1827–1833. doi: 10.1105/tpc.114.135541

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

Emms, D. M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16:157. doi: 10.1186/s13059-015-0721-2

Emms, D. M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20:238. doi: 10.1186/s13059-019-1832-y

Emms, D. M., and Kelly, S. (2021). SHOOT: phylogenetic gene search and ortholog inference. *bioRxiv* [Preprint]. doi: 10.1101/2021.09.01.458564

Frangedakis, E., Shimamura, M., Villarreal, J. C., Li, F. -W., Tomaselli, M., Waller, M., et al. (2020). The hornworts: morphology, evolution and development. *New Phytol.* 229, 735–754. doi: 10.1111/nph.16874

Gould, S. B., Waller, R. F., and McFadden, G. I. (2008). Plastid evolution. *Annu. Rev. Plant Biol.* 59, 491–517. doi: 10.1146/annurev.arplant.59.032607.092915

Grosche, C., and Rensing, S. A. (2017). Three rings for the evolution of plastid shape: a tale of land plant FtsZ. *Protoplasma* 254, 1879–1885. doi: 10.1007/s00709-017-1096-x

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., et al. (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

Harris, B. J., Clark, J. W., Schrempf, D., Szőllosi, G. J., Donoghue, P., Hetherington, A. M., et al. (2021). Divergent evolutionary trajectories of bryophytes and tracheophytes from a complex common ancestor of land plants. *bioRxiv* [Preprint]. doi: 10.1101/2021.10.28.466308

Harris, B. J., Harrison, C. J., Hetherington, A. M., and Williams, T. A. (2020). Phylogenomic Evidence for the Monophyly of Bryophytes and the Reductive Evolution of Stomata. *Curr. Biol.* 30, 2001–2012.e2. doi: 10.1016/j.cub.2020.03.048

Hirano, T., Tanidokoro, K., Shimizu, Y., Kawarabayasi, Y., Ohshima, T., Sato, M., et al. (2016). Moss chloroplasts are surrounded by a peptidoglycan wall containing D-amino acids. *Plant Cell* 28, 1521–1532. doi: 10.1105/tpc.16.00104

Jarvis, P., and López-Juez, E. (2013). Biogenesis and homeostasis of chloroplasts and other plastids. *Nat. Rev. Mol. Cell Biol.* 14, 787–802. doi: 10.1038/nrm3702

Kalanon, M., and McFadden, G. I. (2008). The chloroplast protein translocation complexes of *Chlamydomonas reinhardtii*: a bioinformatic comparison of Toc and Tic components in plants, green algae and red algae. *Genetics* 179, 95–112. doi: 10.1534/genetics.107.085704

Knopp, M., Garg, S. G., Handrich, M., and Gould, S. B. (2020). Major Changes in Plastid Protein Import and the Origin of the Chloroplastida. *iScience* 23:100896. doi: 10.1016/j.isci.2020.100896

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

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., et al. (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* 40, D1202–D1210. doi: 10.1093/nar/gkr1090

Leebens-Mack, J. H. M. S., Barker, E. J., Carpenter, M. K., Deyholos, M. A., Gitzendanner, S. W., Graham, I., et al. (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574, 679–685. doi: 10.1038/s41586-019-1693-2

Lefort, V., Longueville, J.-E., and Gascuel, O. (2017). SMS: smart Model Selection in PhyML. *Mol. Biol. Evol.* 34, 2422–2424. doi: 10.1093/molbev/msx149

Li, F.-W., Nishiyama, T., Waller, M., Frangedakis, E., Keller, J., Li, Z., et al. (2020). Anthoceros genomes illuminate the origin of land plants and the unique biology of hornworts. *Nat. Plants* 6, 259–272. doi: 10.1038/s41477-020-0618-2

Li, F.-W., Villarreal, A. J., and Szövényi, P. (2017). Hornworts: an Overlooked Window into Carbon-Concentrating Mechanisms. *Trends Plant Sci.* 22, 275–277. doi: 10.1016/j.tplants.2017.02.002

Ling, Q., Huang, W., Baldwin, A., and Jarvis, P. (2012). Chloroplast biogenesis is regulated by direct action of the ubiquitin-proteasome system. *Science* 338, 655–659. doi: 10.1126/science.1225053

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., and Zdobnov, E. M. (2021). BUSCO Update: novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol. Biol. Evol.* 38, 4647–4654. doi: 10.1093/molbev/msab199

Martin, A., Lang, D., Hanke, S. T., Mueller, S. J., Sarnighausen, E., Vervliet-Scheebaum, M., et al. (2009). Targeted gene knockouts reveal overlapping functions of the five physcomitrella patens ftsz isoforms in chloroplast division, chloroplast shaping, cell patterning, plant development, and gravity sensing. *Mol. Plant* 2, 1359–1372. doi: 10.1093/mp/ssp076

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., et al. (2020). IQ-TREE 2: new Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* 37, 1530–1534. doi: 10.1093/molbev/msaa015

Nakai, M. (2015a). The TIC complex uncovered: the alternative view on the molecular mechanism of protein translocation across the inner envelope membrane of chloroplasts. *Biochim. Biophys. Acta* 1847, 957–967. doi: 10.1016/j.bbabio.2015.02.011

Nakai, M. (2015b). Ycf1: a green TIC: response to the de Vries et al. Commentary. *Plant Cell* 27, 1834–1838. doi: 10.1105/tpc.15.00363

Osteryoung, K. W., and Pyke, K. A. (2014). Division and dynamic morphology of plastids. *Annu. Rev. Plant Biol.* 65, 443–472. doi: 10.1146/annurev-arplant-050213-035748

Ouyang, M., Li, X., Ma, J., Chi, W., Xiao, J., Zou, M., et al. (2011). LTD is a protein required for sorting light-harvesting chlorophyll-binding proteins to the chloroplast SRP pathway. *Nat. Commun.* 2:277. doi: 10.1038/ncomms1278

Ouyang, M., Li, X., Zhang, J., Feng, P., Pu, H., Kong, L., et al. (2020). Liquid-Liquid Phase Transition Drives Intra-chloroplast Cargo Sorting. *Cell* 180, 1144–1159.e20. doi: 10.1016/j.cell.2020.02.045

Park, E., Ménétret, J.-F., Gumbart, J. C., Ludtke, S. J., Li, W., Whynot, A., et al. (2014). Structure of the SecY channel during initiation of protein translocation. *Nature* 506, 102–106. doi: 10.1038/nature12720

Pyke, K. A., and Leech, R. M. (1992). Chloroplast division and expansion is radically altered by nuclear mutations in Arabidopsis thaliana. *Plant Physiol.* 99, 1005–1008. doi: 10.1104/pp.99.3.1005

Raven, J. A., and Edwards, D. (2014). Photosynthesis in Bryophytes and Early Land Plants. *Diversif. Evol. Environ.* 37, 29–58. doi: 10.1007/978-94-007-6988-5

Rensing, S. A., Lang, D., Zimmer, A. D., Terry, A., Salamov, A., Shapiro, H., et al. (2008). The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science* 319, 64–69. doi: 10.1126/science.1150646

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

Richardson, L. G. L., and Schnell, D. J. (2020). Origins, function, and regulation of the TOC–TIC general protein import machinery of plastids. *J. Exp. Bot.* 71, 1226–1238. doi: 10.1093/jxb/erz517

Vaughn, K. C., Ligrone, R., Owen, H. A., Hasegawa, J., Campbell, E. O., Renzaglia, K. S., et al. (1992). The anthocerote chloroplast: a review. *New Phytol.* 120, 169–190. doi: 10.1111/j.1469-8137.1992.tb05653.x

Villarreal, J. C., Forrest, L. L., Wickett, N., and Goffinet, B. (2013). The plastid genome of the hornwort Nothoceros aenigmaticus (*Dendrocerotaceae*): phylogenetic signal in inverted repeat expansion, pseudogenization, and intron gain. *Am. J. Bot.* 100, 467–477. doi: 10.3732/ajb.1200429

Villarreal, J. C., and Renner, S. S. (2012). Hornwort pyrenoids, carbon-concentrating structures, evolved and were lost at least five times during the last 100 million years. *Proc. Natl. Acad. Sci. U. S. A.* 109, 18873–18878. doi: 10.1073/pnas.1213498109

Xu, X., Ouyang, M., Lu, D., Zheng, C., and Zhang, L. (2021). Protein Sorting within Chloroplasts. *Trends Cell Biol.* 31, 9–16. doi: 10.1016/j.tcb.2020.09.011

Zhang, J., Fu, X. X., Li, R. Q., Zhao, X., Liu, Y., Li, M. H., et al. (2020). The hornwort genome and early land plant evolution. *Nat. Plants* 6, 107–118. doi: 10.1038/s41477-019-0588-4

Zhang, M., Schmitz, A. J., Kadirjan-Kalbach, D. K., TerBush, A. D., and Osteryoung, K. W. (2013). Chloroplast division protein ARC3 regulates chloroplast FtsZ-ring assembly and positioning in Arabidopsis through interaction with FtsZ2. *Plant Cell* 25, 1787–1802. doi: 10.1105/tpc.113.111047

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

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

Copyright © 2022 MacLeod, Raval, Stockhorst, Knopp, Frangedakis and Gould. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 01 April 2022
doi: 10.3389/fpls.2022.842260

Check for updates

OPEN ACCESS

Edited by:
Gerald Matthias Schneeweiss, University of Vienna, Austria

Reviewed by:
František Zedek, Masaryk University, Czechia
Ludmila Cristina Oliveira, Academy of Sciences of the Czech Republic (ASCR), Czechia

***Correspondence:**
Miguel A. García
mgarcia@rjb.csic.es
Andrea Pedrosa-Harand
andrea.harand@ufpe.br

†ORCID:
Amalia Ibiapino
orcid.org/0000-0002-2613-5259
Miguel A. García
orcid.org/0000-0002-0366-043X
Bruno Amorim
orcid.org/0000-0002-8109-9254
Mariana Baez
orcid.org/0000-0002-7874-6385
Mihai Costea
orcid.org/0000-0003-3049-1763
Saša Stefanović
orcid.org/0000-0001-8290-895X
Andrea Pedrosa-Harand
orcid.org/0000-0001-5213-4770

‡*These authors have contributed equally to this work and share first authorship*

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

Received: *23 December 2021*
Accepted: *03 March 2022*
Published: *01 April 2022*

Citation:
Ibiapino A, García MA, Amorim B, Baez M, Costea M, Stefanović S and Pedrosa-Harand A (2022) The Evolution of Cytogenetic Traits in Cuscuta (Convolvulaceae), the Genus With the Most Diverse Chromosomes in Angiosperms. Front. Plant Sci. 13:842260. doi: 10.3389/fpls.2022.842260

The Evolution of Cytogenetic Traits in *Cuscuta* (Convolvulaceae), the Genus With the Most Diverse Chromosomes in Angiosperms

Amalia Ibiapino[1†‡], Miguel A. García[2*†‡], Bruno Amorim[3†], Mariana Baez[4†], Mihai Costea[5†], Saša Stefanović[6†] and Andrea Pedrosa-Harand[1*†]

[1]Laboratory of Plant Cytogenetics and Evolution, Department of Botany, Federal University of Pernambuco, Recife, Brazil, [2]Real Jardín Botánico-CSIC, Madrid, Spain, [3]Postgraduate Program of Biotechnology and Natural Resources of the Amazonia (PPGMBT), State University of Amazonas, Manaus, Brazil, [4]Plant Breeding Department, University of Bonn, Bonn, Germany, [5]Department of Biology, University of Wilfrid Laurier, Waterloo, ON, Canada, [6]Department of Biology, University of Toronto Mississauga, Mississauga, ON, Canada

Karyotypes are characterized by traits such as chromosome number, which can change through whole-genome duplication and dysploidy. In the parasitic plant genus *Cuscuta* (Convolvulaceae), chromosome numbers vary more than 18-fold. In addition, species of this group show the highest diversity in terms of genome size among angiosperms, as well as a wide variation in the number and distribution of 5S and 35S ribosomal DNA (rDNA) sites. To understand its karyotypic evolution, ancestral character state reconstructions were performed for chromosome number, genome size, and position of 5S and 35S rDNA sites. Previous cytogenetic data were reviewed and complemented with original chromosome counts, genome size estimates, and rDNA distribution assessed *via* fluorescence *in situ* hybridization (FISH), for two, seven, and 10 species, respectively. Starting from an ancestral chromosome number of $x = 15$, duplications were inferred as the prevalent evolutionary process. However, in holocentric clade (subgenus *Cuscuta*), dysploidy was identified as the main evolutionary mechanism, typical of holocentric karyotypes. The ancestral genome size of *Cuscuta* was inferred as approximately 1C = 12 Gbp, with an average genome size of 1C = 2.8 Gbp. This indicates an expansion of the genome size relative to other Convolvulaceae, which may be linked to the parasitic lifestyle of *Cuscuta*. Finally, the position of rDNA sites varied mostly in species with multiple sites in the same karyotype. This feature may be related to the amplification of rDNA sites in association to other repeats present in the heterochromatin. The data suggest that different mechanisms acted in different subgenera, generating the exceptional diversity of karyotypes in *Cuscuta*.

Keywords: character evolution, ancestral chromosome number, genome size, ribosomal DNA, heterochromatin, karyotype evolution

INTRODUCTION

Eukaryotes vary in their chromosome constitution and are often characterized by their karyotypes, including both chromosome number and morphology. Among flowering plants, chromosome number has a wide range of variation from $2n = 4$ to $2n = 640$ (Uhl, 1978; Roberto, 2005). The distribution of chromosome numbers in any given monophyletic group allows the identification of one or more chromosome numbers that are considered the ancestral haploid number or basic number of each clade, referred to as x (Guerra, 2008; Mayrose and Lysak, 2021).

From an evolutionary perspective, changes in chromosome number can occur through several mechanisms among which whole-genome duplications within a lineage or autopolyploidy is of special importance (Heslop-Harrison and Schwarzacher, 2011; Alix et al., 2017; Mayrose and Lysak, 2021). Polyploidy can also result from a hybridization event involving two different lineages, a process through which allopolyploids are established (Qiu et al., 2020). Another important source of changes involves ascending and descending dysploidy, that is, a stepwise gain and loss of chromosomes due to structural rearrangements. Descending dysploidy results from incorrect double-stand break repair in two or more chromosomes resulting in chromosome fusion by translocation. This fused chromosome can be inherited by the offspring. In monocentric chromosomes, fusion is usually followed by the elimination or inactivation of one of the centromeres (Guerra, 2008; Schubert and Lysak, 2011; Carta et al., 2020; Mayrose and Lysak, 2021). Centric fission is considered the most common type of ascending dysploidy. The break within a centromere or the wrong centromeric division gives rise to two chromosomes that will be inherited if their function is not impaired (Mayrose and Lysak, 2021). While dysploidy increases or decreases the number of chromosomes mostly preserving the genetic content, aneuploidy is the addition or deletion of one or more chromosomes. Aneuploids can originate in a variety of ways, with mis-segregation during meiosis or mitosis being the most common cause (e.g., Mandáková and Lysak, 2018). The establishment of aneuploids is considered to be uncommon because of imbalance in gene dosage, irregular meiosis, and loss of fertility (Mayrose and Lysak, 2021). Molecular phylogenies have contributed not only to estimate the ancestral chromosome number of a particular clade, but also to our understanding of the polarity of chromosome changes.

In addition to the chromosome number, the evolution of different karyotype features, such as chromosomal bands, number and distribution of ribosomal DNA (rDNA) sites, and genome size can be understood in the light of evolution within a clade and can be correlated, among others, to species diversification (Vaio et al., 2013; Costa et al., 2017; García et al., 2017; Sader et al., 2019). These analyses are performed by mapping and comparing cytogenetic data within phylogenetic trees. The integration of phylogenetic and cytogenetic data also enables reconstructing the ancestral states of cytogenetic characters, evaluating different scenarios of trait evolution. Methods based on parsimony or, more commonly, on probabilistic models have allowed to test chromosome evolution hypotheses within a phylogenetic context, determining characters such as ancestral chromosome numbers (Revell, 2012; Glick and Mayrose, 2014; Maddison and Maddison, 2018; Rice and Mayrose, 2021). Tools such as ChromEvol can estimate the ancestral chromosome number along each branch of a phylogeny while also inferring events like polyploidy and dysploidy (Glick and Mayrose, 2014). Other analytical tools, for example, the R package phytools, allow the reconstruction of ancestral genome size (Revell, 2012), or Mesquite, which was used to reconstruct ancestral states of any characters, including the number and position of heterochromatic bands, as well as 5S and 35S rDNA sites (Revell, 2012; Glick and Mayrose, 2014; Maddison and Maddison, 2018). These approaches are particularly relevant when dealing with extensive samplings or highly variable groups. More recently, Yoshida and Kitano (2021) have proposed a probabilistic method of karyotype evolution incorporating both chromosome and arm numbers, whereby they allowed for a consideration of chromosome morphology as well.

Cuscuta L. (dodders; Convolvulaceae) is a cytogenetically highly diverse genus, with chromosome numbers ranging from $2n = 8$ to $2n = 150$ (Pazy and Plitmann, 1995; García and Castroviejo, 2003; García et al., 2019). The basic numbers for the genus were proposed to be $x = 15$ and $x = 7$ (Fogelberg, 1938; García and Castroviejo, 2003). Most species are diploids with $2n = 30$, but also allopolyploid and autopolyploid species have been documented (García et al., 2014, 2018). Furthermore, among the approximately 200 species of *Cuscuta* (Costea et al., 2015a) genome size varies more than 128-fold, from 1C = 0.27 Gbp in *C. australis* R.Br. to 1C = 34.73 Gbp in *C. reflexa* Roxb. (Sun et al., 2018; Neumann et al., 2020), the highest variation documented for a single genus in angiosperms. This genus is divided into four subgenera, each one with particular cytogenetic features, suggesting strong phylogenetic signals for cytogenetic characters in the group (García et al., 2019; Ibiapino et al., 2022). Subgenus *Cuscuta* is characterized by the presence of holocentric chromosomes; subgen. *Grammica* (Lour.) Yunck. shows the largest variation in chromosome size and number as well as genome size, with confirmed cases of auto- and allopolyploidy; subgen. *Monogynella* (Des Moul.) Peter, Engl. & Prantl includes species with the largest genomes and chromosomes (Fogelberg, 1938; Pazy and Plitmann, 1994; García and Castroviejo, 2003; Guerra and García, 2004; McNeal et al., 2007; García et al., 2019; Ibiapino et al., 2019; Neumann et al., 2020); and finally subgen. *Pachystigma* (Engelm.) Baker & C. H. Wright comprises species with conspicuously bimodal karyotypes. Intraspecific chromosome number variation has been also reported. In species such as *Cuscuta epithymum* (L.) L. and *Cuscuta planiflora* Ten., chromosome number can differ among populations. This variation is even more intriguing in *C. epithymum*, which has holocentric chromosomes and shows $2n = 14$, 16, 28, 30, 32, and 34 in different populations (García and Castroviejo, 2003).

Given the currently available evidence, *Cuscuta* is the genus with the broadest chromosome diversity of all angiosperms.

Abbreviations: AIC, Akaike Information Criterion; CMA, Chromomycin A3; DAPI, 4,6-Diamidino-2-phenylindole; FISH, Fluorescence *in situ* hybridization; PCR, Polymerase chain reaction; rDNA, Ribosomal DNA; BBM, Bayesian Binary MCMC; RASP, Reconstruct Ancestral State in Phylogenies.

No other genus has both holocentric and monocentric chromosomes as well as such a diversity in chromosome size and numbers together with up to a 128-fold difference in genome size. Furthermore, this enormous variation is found at a very low (species) phylogenetic level, which makes this lineage a very tractable system to study genome evolution. Only the carnivorous clade of Caryophyllales shows similar karyotypic diversity but lower differences in genome size: holocentric chromosomes in Droseraceae (1C = 0.24–5.46 Gbp), small monocentric chromosomes in Nepenthaceae (1C = 0.67–1.36 Gbp), and big monocentric chromosomes in Drosophyllaceae (1C = 10.42 Gbp; Veleba et al., 2017, 2020). *Cuscuta* is remarkable because it shows more variation in chromosome and genome size than the five families of carnivorous Caryophyllales even though the clade age of the latter is estimated in the late Cretaceous, c. 84 Mya (Biswal et al., 2018), whereas the *Ipomoea-Cuscuta* lineages split c. 33 Mya (Sun et al., 2018) and the crown age of *Cuscuta* are estimated at 23.0–20.5 Mya (Neumann et al., 2020).

Other parameters, such as the number and position of heterochromatic bands and 5S and 35S rDNA sites, have been comparatively less studied. Nevertheless, the few species investigated still revealed an enormous variation. *Cuscuta denticulata* Engelm. showed one pair of CMA^+/$DAPI^+$ bands, one pair of 5S rDNA, and one pair of 35S, while *C. monogyna* Vahl presented at least 90 CMA^+ bands, 80 $DAPI^+$ bands, 36 5S rDNA sites, and 30 35S rDNA sites (Ibiapino et al., 2019, 2020).

Taken together, this striking karyotypic variation combined with a well-resolved phylogeny (García et al., 2014) makes *Cuscuta* an excellent model for studying karyotypic evolution events in flowering plants. Therefore, the aim of this work was to reconstruct the ancestral states for characters such as chromosome number, genome size, and the position of ribosomal DNA sites in the genus *Cuscuta*. To this end, we reviewed all available data and expanded the banding and rDNA distribution data for 10 previously unstudied species from different clades, six new genome size estimates, and two new chromosome counts, to understand how karyotype evolution occurred and to infer the main events involved in these changes within each subgenus and among subgenera. We also provide a comparative overview of the evolution of genome size and its relationship to the parasitic lifestyle of the genus in a phylogenetic framework.

MATERIALS AND METHODS

Sequence Sampling and Phylogenetic Analysis

For the phylogenetic reconstruction of ancestral chromosome numbers, we sampled 58 taxa of 57 species of *Cuscuta*, including *C. indecora* Choisy var. *indecora* and *C. indecora* var. *neuropetala* (Engelm.) Hitch. The subgenera *Cuscuta*, *Grammica*, *Monogynella,* and *Pachystigma* were represented by eight, 42, four, and three species, respectively. While the monophyly of *Cuscuta* was never seriously challenged, its outgroup relationships and the phylogenetic position within Convolvulaceae remain unresolved (Stefanović and Olmstead, 2004). For this reason, the interpretation of character evolution was based on the ingroup distribution of character states similar to other character evolution studies conducted recently in the genus (e.g., Ho and Costea, 2018, references therein). We used a total of 226 sequences of nuclear (nrITS and 26S) and plastid markers (*rbc*L and *trn*L-*trn*F) obtained by Stefanović and Costea (2008) and García et al. (2014) deposited in GenBank database (Benson et al., 2012; **Supplementary Table 1**). In addition, new ITS and *trn*L-*trn*F sequences were obtained for *C. globosa* Ridl., because this hexaploid was not included in previous phylogenetic works. The methods of DNA extraction, amplification, and sequencing were those detailed in Stefanović et al. (2007). Sequences were uploaded to GenBank with accession numbers OL362011 (ITS) and OL362010 (*trn*L-*trn*F). To align the sequences, the plugin MUSCLE was used in the Geneious v. 7.1.9 software (Kearse et al., 2012).

The phylogenetic relationships were reconstructed using Bayesian Inference (BI) analysis. jModelTest v.2.1.6 (Darriba et al., 2012) selected GTR + I + gamma as the best model of DNA substitution for all analyzed regions, except for *trn*L-*trn*F, which had GTR + gamma as best model. We used MrBayes v. 3.2.6. (Ronquist et al., 2012) to perform BI using the concatenated sequences and selected models with two independent runs with four Markov Chain Monte Carlo (MCMC), sampling every 1,000 generations in a total of 15,000,000 generations. Both BI runs were evaluated in Tracer v.1.6 (Rambaut et al., 2014) to verify if the estimated sample sizes (ESS) for each parameter were higher than 200. The consensus tree was generated in MrBayes with a burn-in of 25%. The consensus tree with the posterior probability (PP) was visualized and edited in FigTree v. 1.4.2. (Rambaut, 2014). The jModelTest and BI analysis were performed through the CIPRES Science Gateway (Miller et al., 2010).

Slide Preparation and FISH

New data on the number and position of rDNA sites for 10 species of *Cuscuta* were obtained for this study (**Table 1**, voucher information in García et al., 2019). Young shoot tips or flower buds were used for slide preparation according to Ibiapino et al. (2019). Double CMA/DAPI staining was performed as described in Ibiapino et al. (2022). The images were captured with a COHU CCD camera attached to a Leica DMLB fluorescence microscope equipped with Leica QFISH software. After image capture, slides were destained for 30 min in Carnoy and 1 h in absolute ethanol and stored for *in situ* hybridization at −20°C. The destained slides were subjected to fluorescent *in situ* hybridization (FISH) according to the protocol described in Pedrosa et al. (2002). Two rDNA probes were used as: the PCR amplified insert of D2 from *Lotus japonicus* (Regel) K. Larsen (5S rDNA; Pedrosa et al., 2002) and p*Ta*71 from wheat (25-28S, 5.8S, and 18S rDNA; Gerlach and Bedbrook, 1979). Probes were labeled by nick translation with Cy3-dUTP (5S) and digoxigenin 11-dUTP (35S). The 5S was labeled in a reaction with total volume of 12.5 µl containing 1 µg of PCR amplified

TABLE 1 | Data of genome size, 5S and 35S ribosomal DNA sites number, and position in species of the genus *Cuscuta*.

Species	1C (Gbp)	5S/35S*/**	References (Genome size/rDNA)
Cuscuta americana	0.68 and 0.69		Neumann et al., 2020, this study
Cuscuta approximata		2T + 4I/2T	Guerra and García, 2004
Cuscuta australis	0.27, 0.34 and 0.69	2I/2P	Sun et al., 2018, Neumann et al., 2020, this study/This study
Cuscuta californica	0.39		Neumann et al., 2020
Cuscuta campestris	0.45, 0.55 and 0.58	4I/2I + 2P	This study, Neumann et al., 2020, Vogel et al., 2018/This study
Cuscuta cephalanthi	3.68 and 3.83		This study, McNeal et al., 2007
Cuscuta chilensis	2.80		McNeal et al., 2007
Cuscuta compacta	3.24 and 7.67		This study, McNeal et al., 2007
Cuscuta denticulata		2I/2P	Ibiapino et al., 2019
Cuscuta epilinum	1.54 and 3.38		McNeal et al., 2007; Neumann et al., 2020
Cuscuta epithymum	0.53 (2*n* = 14)	4I/2T	Neumann et al., 2020/This study
Cuscuta exaltata	20.51		McNeal et al., 2007
Cuscuta europaea	1.05 and 1.17		McNeal et al., 2007; Neumann et al., 2020
Cuscuta globosa	1.79	6I/2I + 2P	This study/ This study
Cuscuta glomerata	5.16	2I/2P	This study/This study
Cuscuta gronovii	3.58		Neumann et al., 2020
Cuscuta gronovii (C PA)	6.75		McNeal et al., 2007
Cuscuta gronovii (NJ)	3.70		McNeal et al., 2007
Cuscuta gronovii (OH)	3.51		McNeal et al., 2007
Cuscuta gronovii (SE PA)	2.14		McNeal et al., 2007
Cuscuta howelliana		2I/2P	This study
Cuscuta indecora	22.68, 24.46 and 32.05	6I + 4I/4P	Ibiapino et al., 2019; McNeal et al., 2007/Ibiapino et al., 2020 Neumann et al., 2020
Cuscuta japonica	25.58		Neumann et al., 2020
Cuscuta lupuliformis	21.97		McNeal et al., 2007
Cuscuta monogyna	32.45 and 33.05	36/30	Ibiapino et al., 2020/Ibiapino et al., 2020; Neumann et al., 2020
Cuscuta nevadensis		6I/8I + 2P	Ibiapino et al., 2019
Cuscuta nitida		2I/4P	Ibiapino et al., 2022
Cuscuta obtusiflora	0.77		McNeal et al., 2007
Cuscuta partita	1.83	2I/2P	This study/This study
Cuscuta pentagona	0.55 and 0.57		McNeal et al., 2007; Neumann et al., 2020
Cuscuta polygonorum	0.79		McNeal et al., 2007
Cuscuta psorothamnensis		6I/2I + 2P	This study
Cuscuta purpurata	2.96		This study
Cuscuta racemosa	1.39	4I/2I + 2P	This study/This study
Cuscuta reflexa	34.73		Neumann et al., 2020
Cuscuta rostrata	3.98		McNeal et al., 2007
Cuscuta sandwichiana	1.80	2I/2I + 2P	This study/This study
Cuscuta veatchii	2.85	6I/2I + 2P	McNeal et al., 2007/ Ibiapino et al., 2019
OUTGROUPS			
Calystegia sepium	0.73		Bai et al., 2012
Calystegia hederacea	1.28		Guo et al., 2015
Convolvulus arvensis	0.65		Pustahija et al., 2013
Convolvulus canariensis	1.01		Suda et al., 2005
Convolvulus cantabricus	1.08		Pustahija et al., 2013
Convolvulus floridus	1.04		Suda et al., 2003
Convolvulus perraudieri	1.04		Suda et al. 2003
Convolvulus scoparius	1.04		Suda et al., 2005
Dichondra repens	1.57		Guo et al., 2015

**T* = *terminal, P* = *peri/centromeric, and I* = *interstitial.*

***Ribosomal DNA sites are represented in number of sites, not in pairs.*

DNA, 1× Nick Translation buffer (0.5 M Tris HCl pH 7.5; 50 mM MgCl2), dNTP mix (0.016 mM each of dATP, dCTP, and dGTP), 0.08 mM Cy3-dUTP or Alexa-dUTP, 7.5 U of DNA Polymerase I, and 0.006 U of DNase I. The mixture was incubated at 15°C for 1 h or longer if needed, until most fragments were under 500 bp, and reactions were stopped using 0.5 M EDTA. The 35S was labeled with the Nick Translation kit (Invitrogen). The images were obtained as previously described.

Flow Cytometry

A total of 11 species had their genome sizes estimated by flow cytometry, six of them here for the first time: *C. glomerata* Choisy, *C. partita* Choisy, *C. purpurata* Phil., *C. racemosa* Mart., *C. sandwichiana* Choisy, and *C. globosa* Ridl. A suspension of nuclei from shoot tips was prepared using WPB buffer (Loureiro et al., 2007). The nuclei were stained using propidium iodide and the amount of nuclear DNA was estimated using the CyFlow SL flow cytometer software (Partec, Görlitz, Germany).

Raphanus sativus L. "Saxa" (1C = 0.53 Gbp), *Solanum lycopersicum* L. "Stupické polní rané" (1C = 0.94 Gbp), *Glycine max* (L.) Merr. "Polanka" (1C = 1.20 Gbp), and *Zea mays* L. "CE-777" (1C = 2.57 Gbp) were used as internal standards (Doležel et al., 2007). The final 2C value was based on three different measurements with 5,000 nuclei each sample, and using the equation "(Sample peak mean/Standard peak) × mean 2C DNA content of internal control (Gbp)" and the software FloMax (Partec) for data processing. The 1C value was obtained by dividing the 2C result by two.

Reconstruction of Ancestral Chromosome Numbers

Data on chromosome numbers are summarized in **Table 2**. Numbers were obtained from the Chromosome Count Database (Rice et al., 2015), to which we contributed numerous counts published in several articles on the cytogenetics of the genus (e.g., García et al., 2019 and references therein). As part of our concerted efforts, these data cover at least one species for all the four subgenera, as well as the 18 sections of subgenera *Cuscuta* and *Grammica* recognized by Costea et al. (2015a). Haploid chromosome numbers were used to infer the basic ancestral numbers for each clade and the genus using ChromEvol v. 2.0 (Glick and Mayrose, 2014). To choose the model that best applies to the data set, the first run was made considering all 10 possible models of the program. Then, the model with the smallest Akaike Information Criterion (AIC) value was selected, and this model was submitted to the model adequacy test for adjustment of each selected model parameter (Rice and Mayrose, 2021). The selected model, BASE_NUM_DUPL, considered the most common chromosome number, that is, the number that appears most frequently in the phylogeny, $n = 15$, and its multiples. The parameters included in this model are the rate of increase of a single chromosome (_gainConstR), the rate of decrease of a single chromosome (_lossConstR), the rate of whole-genome duplications (polyploidy; _duplConstR), rate of transitions per base number (_baseNumberR), and the specified number of chromosomes that characterize a phylogenetic group (_baseNumber), noting that this is not the chromosome number at the root of the phylogeny (Glick and Mayrose, 2014; Rice and Mayrose, 2021).

Due to the numerical chromosome variation reported in *C. epithymum* (subgenus *Cuscuta*, $2n = 14$, 16, 28, 30, 32, and 34), *C. planiflora* (sugenus *Cuscuta*, $2n = 14$, 26, 28, and 34), and *C. denticulata* (subgenus *Grammica*, $2n = 30$ and 60), all the counts found were added to the ChromEvol analysis. First, we made a standard run, using the parameters given by the model adequacy test mentioned above. Then, we executed two more runs, one removing the holocentric clade (subgenus *Cuscuta*) from the analysis to test for the influence of holocentric chromosomes and intraspecific chromosome number variation. Considering the presence of holocentric and monocentric chromosomes in the genus (Fogelberg, 1938; Pazy and Plitmann, 1994), it is possible that different evolutionary models better apply for different clades (Márquez-Corro et al., 2019). In the second additional run, we fixed $n = 15$ to the root, because

TABLE 2 | Haploid chromosome numbers (*n*) considered for character reconstruction.

Species	*n*	References
Cuscuta approximata	14	Guerra and García, 2004
Cuscuta babylonica	4	Pazy and Plitmann, 2002
Cuscuta capitata	10	Mehra and Vasudevan, 1972
Cuscuta epilinum	21	McNeal et al., 2007
Cuscuta epithymum	7, 8, 14, 15, 16, and 17	García and Castroviejo, 2003
Cuscuta europaea	7	García and Castroviejo, 2003
Cuscuta pedicellata	5	Pazy and Plitmann, 1995
Cuscuta planiflora	13 and 14	García and Castroviejo, 2003
Cuscuta americana	15	Neumann et al., 2020
Cuscuta australis	15	García and Castroviejo, 2003
Cuscuta bonafortunae	15	García et al., 2019
Cuscuta brachycalyx	15	García et al., 2019
Cuscuta californica	15	Neumann et al., 2020
Cuscuta campestris	28	García et al., 2019
Cuscuta cephalanthi	30	McNeal et al., 2007
Cuscuta chapalana	15	García et al., 2019
Cuscuta chilensis	15	García et al., 2019
Cuscuta chinensis	14	Aryavand, 1987
Cuscuta compacta	15	García et al., 2019
Cuscuta coryli	15	Fogelberg, 1938
Cuscuta corymbosa var. *grandiflora*	15	García et al., 2019
Cuscuta costaricensis	15	García et al., 2019
Cuscuta cotijana	15, 30	García et al., 2019, this study
Cuscuta cuspidata	15	Pazy and Plitmann, 1995
Cuscuta denticulata	15, 30	García et al., 2018
Cuscuta desmouliniana	15	García et al., 2019
Cuscuta erosa	15	García et al., 2019
Cuscuta globosa	45	This study
Cuscuta glomerata	15	García et al., 2019
Cuscuta grandiflora	15	García et al., 2019
Cuscuta gronovii	30	Love, 1982
Cuscuta howelliana	15	García et al., 2019
Cuscuta indecora	15	Ibiapino et al., 2020
Cuscuta indecora var. *neuropetala*	15	Fogelberg, 1938
Cuscuta nevadensis	15	García et al., 2018
Cuscuta obtusiflora	15	García et al., 2019
Cuscuta occidentalis	15	García et al., 2019
Cuscuta pacifica	15	García et al., 2019
Cuscuta partita	15	This study
Cuscuta pentagona	28	Pazy and Plitmann, 1995
Cuscuta psorothamnensis	30	García et al., 2018
Cuscuta purpurata	15	García et al., 2019
Cuscuta racemosa	30	García et al., 2019
Cuscuta salina	*ca.* 15	Pazy and Plitmann, 1995
Cuscuta sandwichiana	*ca.* 75	García et al., 2019
Cuscuta sidarum	15	García et al., 2019
Cuscuta subinclusa	15	García et al., 2019
Cuscuta tinctoria	19	Pazy and Plitmann, 1995
Cuscuta tinctoria var. *floribunda*	15	García et al., 2019
Cuscuta umbrosa	15	García et al., 2019
Cuscuta veatchii	30	Ibiapino et al., 2019
Cuscuta volcanica	15	García et al., 2019
Cuscuta japonica	15	Neumann et al., 2020
Cuscuta lupuliformis	14	McNeal et al., 2007
Cuscuta monogyna	15	García and Castroviejo, 2003
Cuscuta reflexa	16	Neumann et al., 2020
Cuscuta africana	14	Ibiapino et al., 2022
Cuscuta angulata	15	Ibiapino et al., 2022
Cuscuta nitida	14	Ibiapino et al., 2022

this is the basic number proposed for *Cuscuta* from cytogenetic data (Pazy and Plitmann, 2002) and was supported by the data compilation produced in this work. The results were plotted in R using the ChromEvol functions as described in Cusimano et al. (2012).

An additional reconstruction of the ancestral chromosome number was performed in Mesquite version 2.75, using maximum likelihood (Maddison and Maddison, 2018) to compare results with those originated from ChromEvol. However, in Mesquite, the haploid chromosome numbers were categorized into nine states: $n = 4$ (coded as 0), $n = 5$ (1), $n = 7$ (2), $n = 10$ (3), $n = 13$ (4), $n = 14$ (5), $n = 15$ (6), $n = 16$ (7), $n = 19$ (8), and all polyploids from $n = 21$ to $n = 75$ (9). Additionally, we compared the results to the inference of the ancestral state of this character made along the branches using PastML (Ishikawa et al., 2019).[1] We applied the JOINT (highest likelihood) method. As the model assume only one state per sample, we used for *C. epithymum and C. planiflora* the cytotypes analyzed in the present work, $n = 14$, and for *C. denticulata*, $n = 15$.

Reconstruction of Genome Sizes

Genome size estimations for six species were newly obtained for this paper in addition to new assessments for five species with previously published data. The reconstruction was performed for 28 *Cuscuta* species in total (**Table 1**), three of subgenus *Cuscuta,* 20 of subgenus *Grammica*, and five of subgenus *Monogynella.* This trait was analyzed as a continuous character in the phytools package (Revell, 2012). Thirty-three taxa that lacked GS information were excluded from our original tree using the ape package (Paradis et al., 2004) and we included three additional species with known genome size but unknown chromosome numbers (*C. rostrata* Engelm. & A. Gray, *C. polygonorum* Engelm., and *C. exaltata* Engelm.). Both phytools and ape packages were implemented in R (R Core Team, 2020). For this analysis, we considered the value of 1C in Gbp, and for species with two or more genome sizes published, an average was made between the values. In addition, for comparative purposes, a reconstruction of the genome size was also performed in Mesquite using the implemented maximum parsimony analysis.

To address whether the inclusion of outgroups changed significantly the results of the previous analyses, we performed three additional reconstructions of the ancestral genome size of *Cuscuta*. Each one included as outgroup candidates of Convolvulaceae with genome size data available of the two sister groups resolved by Stefanović and Olmstead (2004) as the closest relatives of *Cuscuta*. One of the analyses used two species of *Calystegia* R.Br. and six of *Convolvulus* L. (Convolvuleae, clade 1), another one *Dichondra repens* J.R.Forst. & G.Forst. (Dichondreae, clade 2), and the third one all of them.

Reconstruction of rDNA Ancestral Positions

The reconstruction of ancestral number and positions of the 5S and 35S rDNA sites were performed using Mesquite version 2.75 (Maddison and Maddison, 2018) on the 18 species for which rDNA information was available, 10 of them newly generated for this paper. Both the number and position of sites were transformed into categorical data (discrete characters): centromeric/pericentromeric position, interstitial, terminal/subterminal, and "mix" (when more than one of the previous conditions occurs in the same karyotype) as proposed by García et al. (2017). For the number of 5S sites, the characters were categorized as 1, 2, 3, 5, or 18 pairs. The number of 35S sites was categorized as 1, 2, 5, or 15 pairs of sites. Ancestral character states were inferred using maximum likelihood (Vaio et al., 2013). Due to the inconclusive results obtained in the rDNA sites number reconstruction using Mesquite, a second reconstruction was conducted using the Bayesian Binary MCMC (BBM) tool (Ali et al., 2012) implemented in the software Reconstruct Ancestral State in Phylogenies—RASP 4.2 (Yu et al., 2015, 2020) using the default parameters.

RESULTS

Phylogenetic Reconstruction

In total, 57 species of *Cuscuta* with DNA sequences and cytogenetic data available were sampled for the phylogenetic reconstruction, representing approximately 30% of the *ca.* 200 known species of the genus. The four subgenera were recovered as monophyletic each. Subgenus *Monogynella* was represented by four out of 15 species (26.67%), *Cuscuta* by eight out of 22 (36.36%), *Pachystigma* by three out of five (60%), and *Grammica* by 42 out of 150 (28.6%). The phylogenetic relationships obtained in this study were consistent with those based on a larger dataset reported by García et al. (2014), in which each of the four subgenera was strongly supported as monophyletic, with subgenus *Monogynella* as sister to the rest. The trees resolved the same relationships between the sections of subgenera *Cuscuta* and *Grammica* as those obtained by García et al. (2014; **Supplementary Figure 1**). The phylogenetic position of *C. globosa* was resolved as a member of section *Gracillimae* (clade N). Based solely on the description of the type, this species had previously been included in section *Racemosae* (clade C; Costea et al., 2015a). Here, it is placed in section *Gracillimae* based on molecular data, in agreement with the morphological features studied in the type and our new collections of this species.

Chromosome Number, rDNA Site, and Genome Size Variation in *Cuscuta*

Most *Cuscuta* species with published chromosome number are diploids with up to $2n = 38$ (40 species plus two varieties, **Table 2**). Another 11 species are polyploids, mostly with $2n = 60$. Of the species included in this study, four of them are known to have diploid and tetraploid populations: *C. planiflora*, *C. epithymum*, *C. cotijana* ($2n = 60$, a new cytotype; **Supplementary Figure 2**), and *C. denticulata*. Two new counts were included in this work, *C. partita* ($2n = 30$) and *C. globosa* ($2n = 90$; **Supplementary Figure 2**). The majority of polyploids belongs to subgenus *Grammica*, with the exception of some

[1]https://pastml.pasteur.fr/

tetraploids and hexaploids of subgenus *Cuscuta*, such as *C. approximata* ($2n = 28$) or *C. epilinum* Weihe ($2n = 42$), considering a lower basic number for this subgenus (see below). The smallest number found was $2n = 8$ in *C. babylonica* Aucher ex Choisy (*Cuscuta*), while the largest number was $2n = 150$ in *C. sandwichiana* (*Grammica*). *Cuscuta epithymum* (*Cuscuta*) presents numerical intraspecific variation with $2n = 14$, 16, 28, 30, 32, and 34.

For seven species, 5S and 35S rDNA site number and location were previously published (**Table 1**; Guerra and García, 2004; Ibiapino et al., 2019, 2020, 2022). New rDNA data were obtained for 10 additional species, *C. australis* ($2n = 30$), *C. campestris* Yunck. ($2n = 56$), *C. epithymum* (cytotype with $2n = 30$), *C. howelliana* P. Rubtzoff ($2n = 30$), *C. partita* ($2n = 30$), *C. psorothamnensis* Stefanović, M. A. García & Costea ($2n = 60$), *C. racemosa* ($2n = 60$), *C. sandwichiana* ($2n = 150$), *C. globosa* ($2n = 90$), and *C. glomerata* ($2n = 30$). All rDNA sites in *Cuscuta* were colocalized with CMA^+ bands. Most species presented at least one pair of $CMA^+/DAPI^-$ bands colocalized with nucleolus organizer regions (NOR) in proximal regions. Interstitial bands, when present, were weaker and smaller. Only in the holocentric *C. epithymum* did these bands occur in the terminal regions and were present in most chromosomes (**Figures 1, 2**). Most species showed only one pair of 5S and one pair of 35S rDNA sites. Usually, 5S sites occurred in interstitial regions, while 35S sites in pericentromeric regions, such as in *C. australis*, *C. howelliana*, *C. partita*, and *C. glomerata* (**Figure 1**). When more than one pair of 5S rDNA was present, these sites were also in interstitial regions, whereas when there were more than one pair of 35S, the extra pairs were interstitial, as in *C. campestris*, *C. racemosa*, *C. psorothamnensis*, *C. globosa,* and *C. sandwichiana* (**Figure 2**), or terminally located, as in *C. epithymum* (**Figure 1**). The largest number of rDNA sites in *Cuscuta* species was observed in *C. monogyna*, with approximately 18 pairs of 5S and 15 pairs of 35S rDNA. Information on the number and distribution of the rDNA in *Cuscuta* can be found in **Table 1**.

As for genome size, *Cuscuta* species varied from 1C = 0.27 Gbp in *C. australis* ($2n = 30$) to 1C = 34.73 Gbp in *C. reflexa* ($2n = 32$), both diploids. This variation represents the lowest and highest genome sizes known for Convolvulaceae (**Figure 3**). Some species showed infraspecific variation in genome size, such as *C. gronovii* Wild. ex Roem. & Schult., with five different values reported, ranging from 1C = 2.14 Gbp to 1C = 6.75 Gbp (**Table 1**).

Ancestral Character State Reconstructions

The chromosome number reconstruction performed in ChromEvol with the best model, BASE_NUM_DUPL, indicated $n = 7$ as the basic ancestral number in *Cuscuta* (**Supplementary Figure 3**). This model considers five parameters, the rates of gains and losses of single chromosomes, duplications, in addition to considering a specific chromosome number that characterizes a phylogenetically close group, and the number variation rate. Based on this model, the variation in chromosome number in *Cuscuta* is most often related to duplication events [with a probability of frequency ($f = 9.2$)], followed by chromosome gains ($f = 8.3$) and losses ($f = 7$). The number $n = 15$ was indicated as ancestor of the subgenera *Grammica* and *Monogynella*. The ancestral number of subgenus *Cuscuta* was $n = 7$, and *Pachystigma* had $n = 14$, with 50% probability, but $n = 7$ was also very likely, with 40% probability (**Supplementary Figure 3**).

The holocentric clade (subgenus *Cuscuta*) includes species with intraspecific numerical variation. To test if its holocentric nature and high chromosome number variation influenced the analysis, we ran ChromEvol without the holocentric clade, following the same parameters described above (**Supplementary Figure 4**). The basic number in this analysis was $n = 15$ ($x = 15$), again with chromosome duplication ($f = 7$), followed by chromosome gains ($f = 5.6$) and chromosome losses ($f = 3.7$) as the main evolutive events. The reconstructed ancestral number for the remaining three subgenera was conserved as $n = 15$ (**Supplementary Figure 4**). Therefore, we repeated ChromEvol analysis fixing $n = 15$ at the base of the genus (**Figure 4**). In this scenario, numerical changes were mainly due to chromosome losses ($f = 14.9$), followed by duplications ($f = 7.6$) and chromosome gains ($f = 6.6$). The reconstructed ancestral numbers for *Grammica* and *Monogynella* were also $n = 15$. In subgenus *Cuscuta*, the basic number was $n = 7$ and the basic number for *Pachystigma* was $n = 14$ (**Figure 4**).

Mesquite and PastML also recovered $n = 15$ as the basic chromosome number for the whole genus. Furthermore, the maximum likelihood analysis suggested $n = 15$ for the subgenera *Monogynella* and *Grammica*. For subgenus *Pachystigma*, the most likely basic number reconstructed was $n = 14$ (with 50% probability), but $n = 15$ was also likely (with 43% probability). Only subgenus *Cuscuta* had three possible basic numbers depending on the analysis. The Mesquite analysis retrieved $n = 7$ and 14 as the most likely, each with approximately 31% probability (**Supplementary Figure 5**). The PastML JOINT approach reconstructed $n = 15$ as the ancestral number for both the entire genus and each of its four subgenera (data not shown).

Considering the smallest genome of 1C = 0.27 Gbp and the largest of 1C = 34.73 Gbp, the reconstructions of ancestral genome size made with phytools suggested that the ancestral genome of *Cuscuta* would be of intermediate size, approximately 20 Gbp without outgroups (**Supplementary Figure 6**), and 1C = 12 Gbp, including available outgroups. The results obtained using *Dichondra* or *Calystegia* and *Convolvulus* or these three genera as outgroup were similar (14, 12, and 12 Gbp, respectively; **Supplementary Figure 6**). Genome size increased in subgenus *Monogynella*, while it decreased in the other subgenera. Only *C. indecora* showed a massive expansion of genome size within subgenus *Grammica* (**Figure 5**), what was also observed with Mesquite (data not shown). Although there is no estimation available for subgenus *Pachystigma,* the long chromosomes of their bimodal karyotypes also suggest an increase in genome size, especially for *C. angulata* (Ibiapino et al., 2022).

The reconstruction of the ancestral number of rDNA sites made by Mesquite was inconclusive (data not shown). The analysis performed using RASP reconstructed 18 pairs of sites

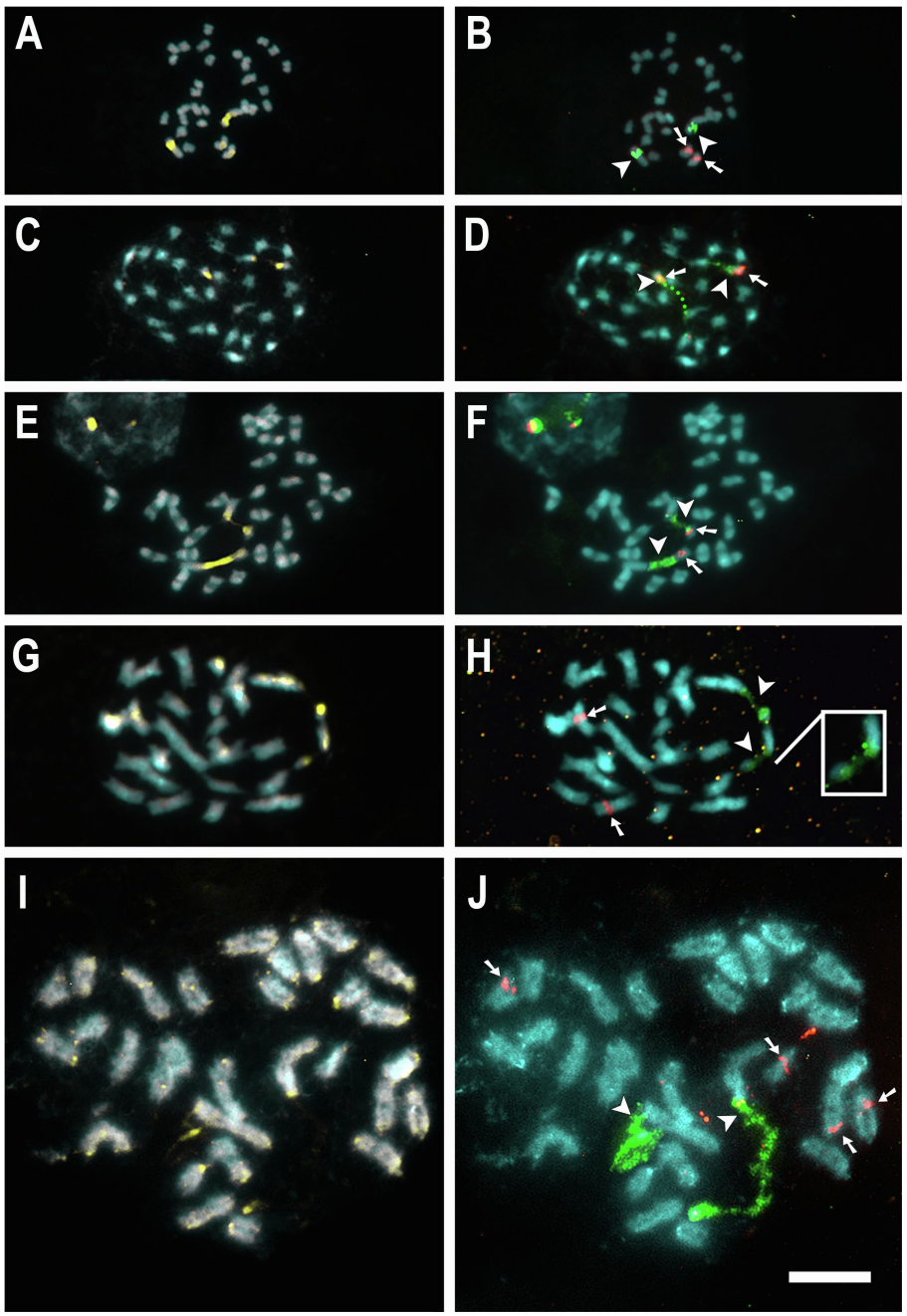

FIGURE 1 | Mitotic metaphases of diploids with karyotypes $2n = 30$. *C. australis* **(A,B)**, *C. howelliana* **(C,D)**, *C. partita* **(E,F)**, *C. glomerata* **(G,H)**, and *C. epithymum* **(I,J)** stained with CMA (yellow) and DAPI (blue) in **A, C, E, G,** and **I**, and with FISH of 5S (red) and 35S (green) rDNA in **B, D, F, H,** and **J**. Arrowheads indicate 35S (green) and arrows indicate 5S (red) rDNA sites. Insets show weak signals in higher contrast. Bar in **J** represents 10 μm; all images at the same magnification.

as ancestral for 5S rDNA and 15 pairs for 35S rDNA, but it is probably due to the presence of numerous sites in *C. monogyna* (**Supplementary Figure 7**). The reconstruction of rDNA site positions indicated that the ancestral position of the 5S rDNA was likely interstitial. The reconstructed position of the 35S rDNA site was inconclusive, with *C. monogyna* showing terminal and interstitial sites. Interstitial 5S rDNA sites were maintained throughout the genus. Only in *C. indecora* additional sites in terminal positions appeared. The 35S rDNA was reconstructed as terminal in the subgenus *Cuscuta* and peri/centromeric in *Grammica*. In this latter subgenus, species with only a pair of 35S sites had them always in peri/centromeric position. When more than one pair of 35S rDNA was present, the extra sites were inferred to have originated in interstitial positions (**Figure 6**).

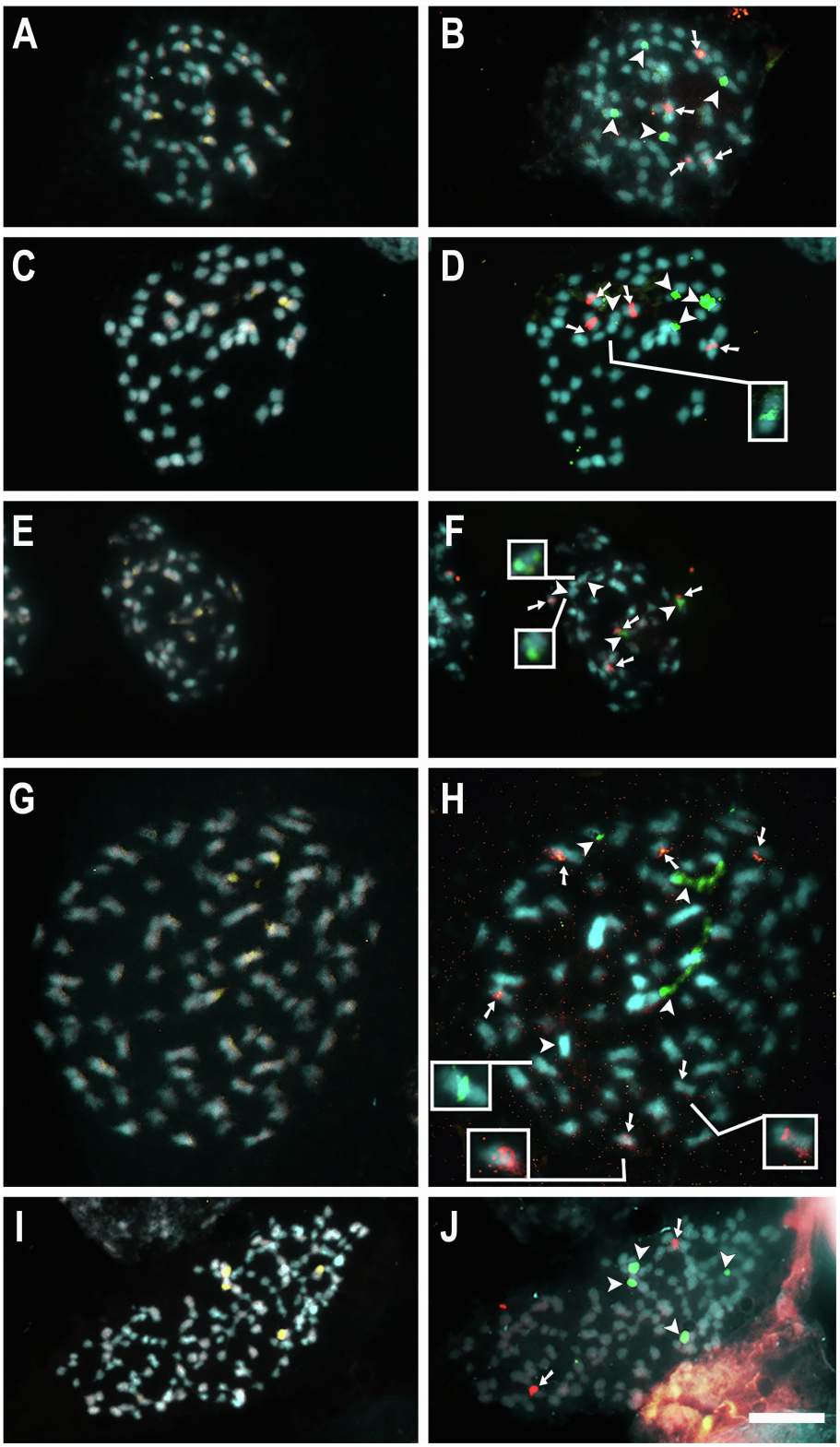

FIGURE 2 | Mitotic metaphases of the polyploids *C. campestris* $2n = 56$ **(A,B)**, *C. racemosa* $2n = 60$ **(C,D)**, *C. psorothamnensis* $2n = 60$ **(E,F)**, *C. globosa* $2n = 90$ **(G,H)**, and *C. sandwichiana* $2n = 150$ **(I,J)** stained with CMA (yellow) and DAPI (blue) in **A, C, E, G,** and **I**, and with FISH of 5S (red) and 35S (green) rDNA in **B, D, F, H,** and **J**. Arrowheads indicate 35S (green) and arrows indicate 5S (red) rDNA sites. Insets show weak signals in higher contrast. Bar in **J** represents 10 μm.

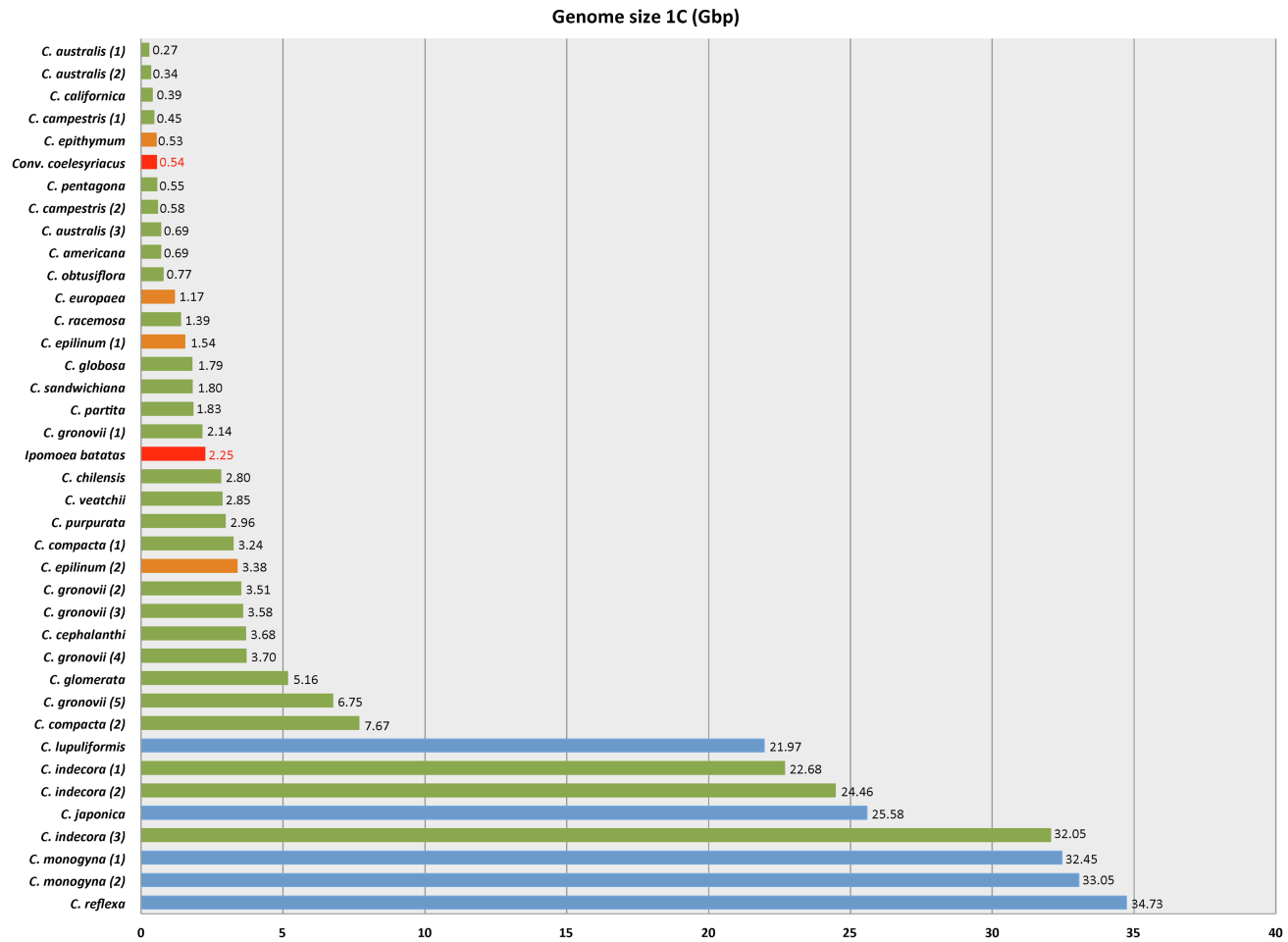

FIGURE 3 | Genome size variation in *Cuscuta* based on the data in **Table 2**. Red bars indicate the lowest and highest known genome sizes in other Convolvulaceae, as reported in the Plant DNA C-values Database (https://cvalues.science.kew.org/). Color bars for *Cuscuta* indicate the subgenera: blue for *Monogynella*, green for *Grammica,* and orange for subgenus *Cuscuta.*

DISCUSSION

Evolution of Chromosome Number in *Cuscuta* and the Uniqueness of the Holocentric Clade

Chromosome number variation across the entire genus *Cuscuta* was almost 19-fold between $2n = 8$ in *C. babylonica* and $2n = 150$ in *C. sandwichiana* (Pazy and Plitmann, 2002; García et al., 2019). The subgenus *Cuscuta* alone showed a variation of over 5-fold (from $2n = 8$ to $2n = 42$). The variation within this subgenus is also associated with intraspecific numerical variation, found in *C. epithymum* ($2n = 14$, 16, 28, 30, 32, and 34) and *C. planiflora* ($2n = 14$, 26, 28, and 34; García and Castroviejo, 2003). Subgenus *Grammica* species diversity was less represented with only ca. 30% of its ca. 150 species (Stefanović et al., 2007; Costea et al., 2015a). However, the variation in chromosome numbers found was similar to that observed in subgenus *Cuscuta*, just over 5-fold ($2n = 28$ to $2n = 150$). In this case, most of the variation is attributable to auto- or allopolyploidy and only a few species have a chromosome number that is not $n = 15$ or a multiple thereof. The additional numbers can be explained by ascending or descending dysploidy. The relative low proportion of species studied suggests that the variation in this subgenus may be underestimated. Hybrid speciation was frequent in subgenus *Grammica* (e.g., Stefanović and Costea, 2008; Costea and Stefanović, 2010; García et al., 2014; Costea et al., 2015b) and a more detailed sampling will probably reveal additional cases of both auto- and allopolyploidy, as was recently reported for the small section *Denticulatae* (clade E; García et al., 2018).

In subgenera *Monogynella* and *Pachystigma*, no large variation in chromosome number was found. In this work, *Monogynella* is represented by five of its 15 species and the most common chromosome numbers found were $2n = 28$, 30, and 32. Two additional species of the subgenus without sequence data were reported to have $2n = 32$ chromosomes, *C. gigantea* Griff. (Khatoon and Ali, 1993) and *C. sharmanum* Mujerjee & P. K. Bhattach. (Mukerjee and Bhattacharya,

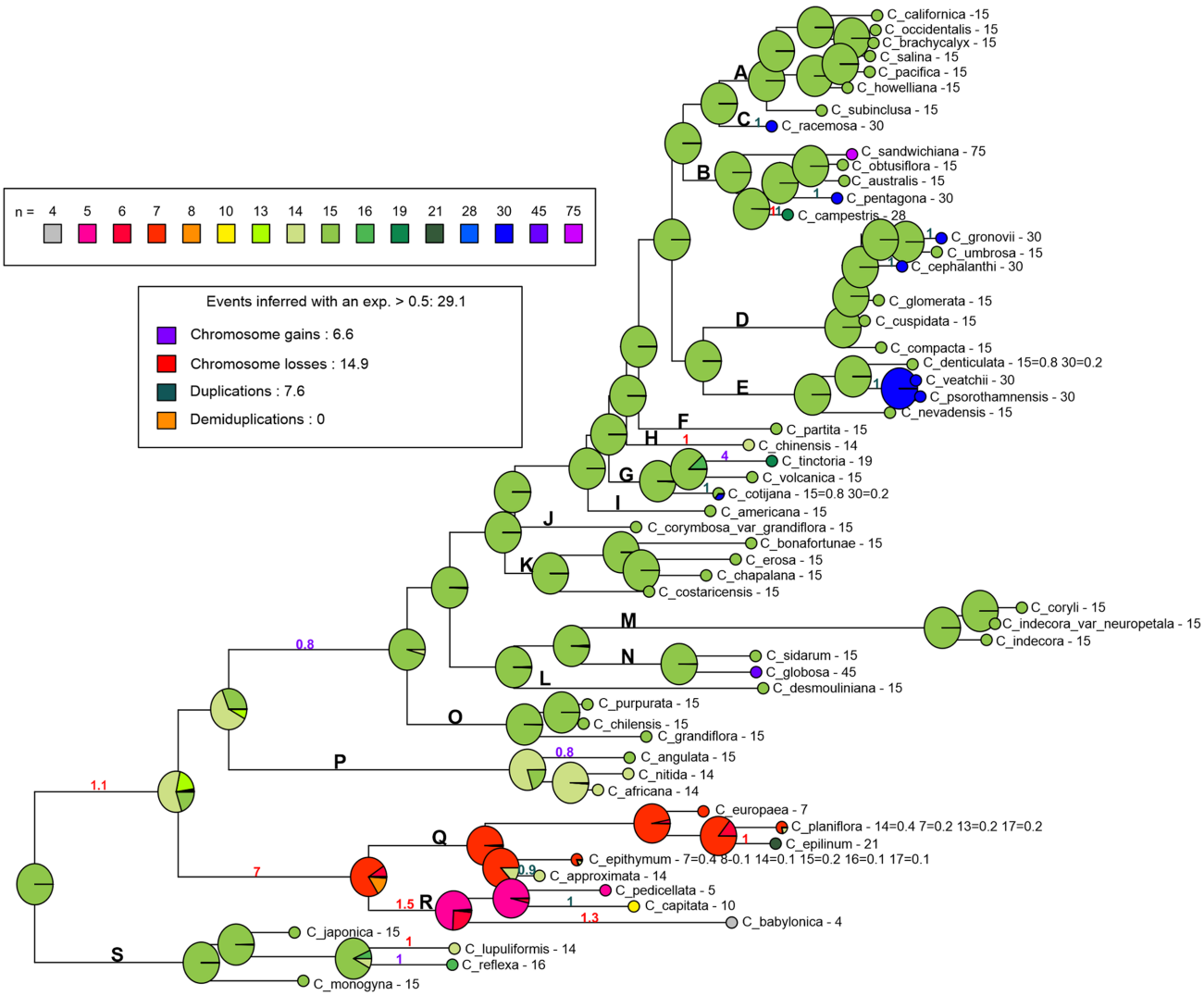

FIGURE 4 | Reconstruction of the chromosome number evolution in *Cuscuta* with the BASE_NUM_DUPL model and the number $n = 15$ fixed in the base. The pie charts at nodes represent the probability of each inferred chromosome number, the numbers along the branches represent the probability of frequencies of the inferred events (gains, losses, duplications, and demiduplications). The bold letters represent the sections described by García et al. (2014): section S (subgenus *Monogynella*), sections R and Q (subgenus *Cuscuta*), section P (subgenus *Pachystigma*), and sections A–O (subgenus *Grammica*).

1970). Only *C. reflexa* has been reported to have a great intraspecific variation of chromosome numbers with triploid, tetraploid, and dysploid or aneuploid individuals (Kaul and Bhan, 1977). The small subgenus *Pachystigma*, represented by three of its total of five species (60%), has $2n = 28$ and 30. The species of this subgenus have bimodal karyotypes; however, both the number of large and small pairs and the distribution of heterochromatic bands are quite variable (Ibiapino et al., 2022).

We used ChromEvol to reconstruct the basic ancestral chromosome number, which indicated the BASE_NUM_DUPL model as the best for our dataset. With this model, the basic ancestral number for *Cuscuta* was $x = 7$, as previously suggested by Fogelberg (1938) and García and Castroviejo (2003). However, most of the chromosome numbers reported in *Cuscuta* are multiples of 15, because 32 species (55%) are diploids with $2n = 30$ and, among polyploids, $2n = 60$ is the most frequent number. The highest numbers reported are $2n = 90$ and 150, which are also multiples of 15. Additional analyses of both Mesquite and Past ML corroborated the number $x = 15$ at the base of the genus *Cuscuta*.

According to the Chromosome Count Data Base (Rice et al., 2015), the most frequent chromosome number in other Convolvulaceae is $n = 15$. Because of the accelerated rates of sequence evolution in *Cuscuta*, the extant sister group of the genus within Convolvulaceae could not be ascertained, but at least two nonparasitic lineages diverged before *Cuscuta* (Stefanović and Olmstead, 2004). Chromosome numbers in these lineages are known for a couple of genera in tribe Cardiochlamyeae: $n = 13$ for *Poranopsis* Roberty and $n = 14$ for *Dinetus* Sweet.

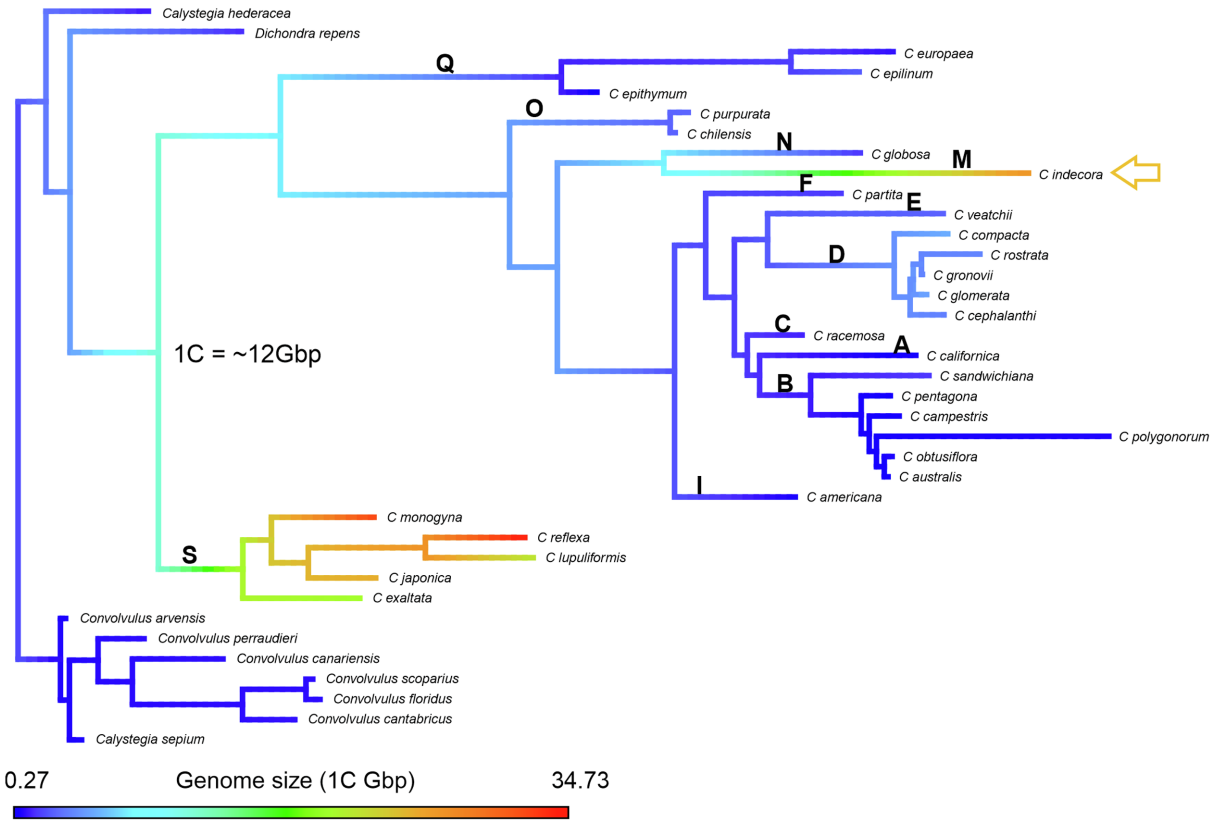

FIGURE 5 | Reconstruction of the genome size evolution *via* Phytools in R. The variation is shown in a color scale: the largest genomes in shades of red and the smallest genomes in shades of blue. The yellow arrow in the reconstruction indicates *C. indecora*, a species of the subgenus *Grammica* in which there was a massive expansion of the genome. The bold letters represent the sections described by García et al. (2014): section S (subgenus *Monogynella*), sections R and Q (subgenus *Cuscuta*), section P (subgenus *Pachystigma*), and sections A–O (subgenus *Grammica*).

Among more closely related lineages chromosome numbers are mostly $n = 14$ or 15, but some genera show greater diversity in chromosome numbers such as *Convolvulus* L. ($n = 9$–30), *Merremia* Endl. ($n = 7$–29), or *Ipomoea* L. ($n = 14$–45). Altogether, it is unlikely that, in *Cuscuta*, the ancestral number has reduced to $x = 7$, followed by independent chromosome duplications and gains in *Monogynella* and *Pachystigma* + *Grammica*. Therefore, we consider $x = 15$ more likely for the genus *Cuscuta*.

One reason for the inference of $x = 7$ for the genus is probably the presence of lower numbers in the subgenus *Cuscuta*. However, this subgenus is exclusively holocentric, and this chromosome type may go through karyotypic changes that are different than those in the other subgenera. Chromosome fusion and fission events can be favored in this karyotype type, as they have a diffuse kinetochore, which facilitates these types of rearrangements (Mandrioli and Manicardi, 2020). In groups where different evolutionary dynamics occur, it is necessary to consider clade specific models, that is, different parts of the phylogeny evolving according to different transition patterns of changes in chromosome numbers (Mayrose and Lysak, 2021). Márquez-Corro et al. (2019) used different methodological approaches to identify diverse patterns of chromosomal evolution in some clades of Cyperaceae. In that case, both a complete tree and subtrees were analyzed, suggesting several evolutionary model transitions in the entire phylogeny of the family. This type of analysis is particularly relevant when applied to the study of clades containing species with holocentric chromosomes, whose karyotypes can exhibit heterogeneous evolution modes. Therefore, we removed the holocentric clade to evaluate the reconstruction. In this analysis, the basic number $x = 15$ was retrieved for both the genus as a whole, and in each of the remaining subgenera, with duplication events as the most frequent resulting in the formation of polyploids. This corroborated the idea that the evolutionary dynamics of holocentrics are indeed different, and it exerted a large influence on the reconstruction of chromosome numbers in *Cuscuta*.

We therefore considered $n = 15$ fixed at the base of the genus *Cuscuta* as the best model to explain the evolution of chromosome numbers of this genus. In this model, seven chromosome losses occurred leading to $x = 7$ after the transition from monocentric to holocentric chromosomes in the subgenus *Cuscuta*. For the holocentric species, the chromosome numbers found were $2n = 8$, 10, 14, 16, 18, 20, 26, 28, 30, 32, 34, and 42. Among the three sections of subgenus *Cuscuta* recognized by Costea et al. (2015a), the lowest chromosome number,

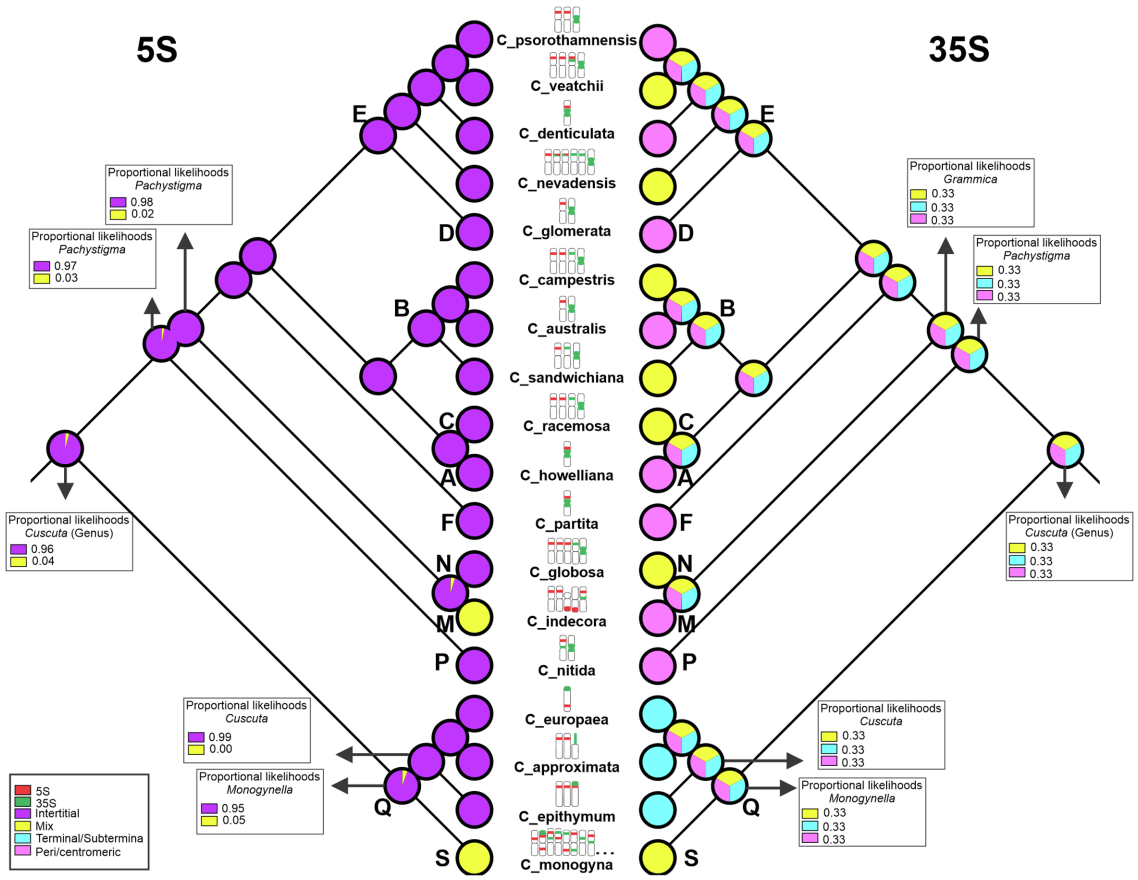

FIGURE 6 | Reconstruction of the position of the 5S (left) and 35S (right) rDNA sites. In subgenus *Monogynella,* there is an increase in the diversity of positions in which the rDNA sites were found, suggesting that the "mix" condition is derived. In *C. monogyna,* not all pairs are represented because it is a species that has more than 30 rDNA sites, but all observed patterns are outlined. The bold letters represent the sections described by García et al. (2014): section S (subgenus *Monogynella*), sections R and Q (subgenus *Cuscuta*), section P (subgenus *Pachystigma*), and sections A–O (subgenus *Grammica*).

$2n = 8$, was found in *C. babylonica*, the only species of the monotypic section *Babylonicae*. This section is generally recovered as sister to section *Epistigma*, a group of five, mostly Asian species with chromosome numbers known in *C. pedicellata* Ledeb. and *C. pulchella* Engelm., both with $2n = 10$, and *C. capitata* Roxb., probably a tetraploid with $2n = 20$. The lineage of sections *Epistigma* and *Babylonicae* may have reduced their chromosome number by descending dysploidy through chromosome fusions. In this lineage, the biggest chromosomes are in *C. babylonica*, which shows the lowest chromosome number.

In the sister lineage, section *Cuscuta* of the subgenus *Cuscuta*, diploid species have at least $2n = 14$ chromosomes, several species such as *C. approximata* and *C. palaestina* Boiss. are tetraploids ($2n = 28$), and *C. epilinum* is a hexaploid ($2n = 42$). Other diploids, not included in our analyses, have higher chromosome numbers, such as $2n = 18$ and 20 reported for *C. nivea* M.A. García (García, 2001). Two species, *C. epithymum* and *C. planiflora*, are known to have diploid and polyploid populations with further variation in chromosome numbers that may have been generated from duplication and ascending dysploidy events. In *C. epithymum* ($2n = 14$, 16, 28, 30, 32, and 34), there are diploid and tetraploid cytotypes with both bimodal and symmetrical karyotypes. Some cytotypes with $2n = 14$ and $2n = 32$ are bimodal, whereas others with $2n = 16$ and $2n = 34$ are symmetrical (García and Castroviejo, 2003; García et al., 2019), suggesting that chromosome fusion or fission events, together with polyploidy, engender this numerical variation. The cytotypes with asymmetrical karyotypes in *C. epithymum* and the other species of subgenus *Cuscuta* have the active NORs located in the longest chromosomes and with hematoxylin staining they are observed associated to the nucleoles. In *C. planiflora* ($2n = 14$, 26, 28, and 34), the populations with $2n = 34$ have an asymmetrical karyotype that together with morphological features indicate that it is an allopolyploid (García, 2001). Both, *C. epithymum* and *C. planiflora* are taxonomically difficult as revealed by the high number of infraspecific taxa that have been described (García and Martín, 2007), and their variation in chromosome numbers and karyotypes may indicate cryptic diversity in these species' complexes.

Genome Size Variation in *Cuscuta* Is Extreme and Is Reflected in Chromosome Sizes

Genome size data for 28 species of *Cuscuta* (**Table 1, Figure 5**) revealed a 128-fold variation between the smallest (1C = 0.27 Gbp in *C. australis*) and the largest genome (1C = 34.73 Gbp in *C. reflexa*). This tremendous variation in *Cuscuta* does not seem to be mainly caused by polyploidy events, despite its high frequency in the genus (García et al., 2019; Ibiapino et al., 2019). Polyploids with higher chromosome numbers, such as *C. globosa* ($2n = 90$) and *C. sandwichiana* ($2n = 150$), both belonging to subgenus *Grammica*, had small genome sizes, 1C = 1.79 Gbp and 1C = 1.8 Gbp, respectively. *Cuscuta*, therefore, fits with the general trend of genome downsizing of polyploids observed in angiosperms (Leitch and Leitch, 2008). Subsequent genome downsizing in polyploids is also known from other parasitic lineages (e.g., in *Orobanche*; Weiss-Schneeweiss et al., 2006).

It is remarkable that the largest genome sizes are found in diploid species such as *C. lupuliformis* Krock. ($2n = 28$) and *C. reflexa* ($2n = 32$), both belonging to subgenus *Monogynella*, with 1C = 21.97 Gbp and 1C = 34.73 Gbp, respectively. Species such as *C. monogyna* (1C = 33.05 Gbp, subgenus *Monogynella*) and *C. indecora* (1C = 24.46 Gbp, subgenus *Grammica*), both diploids, have numerous heterochromatic bands along their chromosomes, all these bands co-localizing with repetitive DNA such as 5S and 35S rDNA or satellite DNAs, indicating that the amplification of repetitive sequences in heterochromatin is involved (Ibiapino et al., 2020; Naumann et al., 2020). However, there is also an accumulation of heterochromatic bands in subgenus *Cuscuta*, but there is no drastic increase in genome size probably because of the reduction in the number of chromosomes associated with the transition to holocentric chromosomes. In the diploid *C. europaea* L., for example, the genome size is, on average, 1C = 1.11 Gbp (McNeal et al., 2007; Neumann et al., 2020). This species has satellite DNAs that occupy a large extension of all 14 chromosomes, which varies in sizes from 2.76 to 6.70 µm, approximately (estimated measurements based on mitotic metaphases reported in Oliveira et al., 2020). In subgenus *Pachystigma*, the differential accumulation of repeats in just a few chromosomes led to the appearance of bimodal karyotypes. *Cuscuta nitida* E. Mey. ex Choisy ($2n = 28$), for example, has an accumulation of different classes of repetitive DNA in only two chromosome pairs, which are, in average, 12.34 and 8.19 µm long, compared to 2.67 µm of the smallest pairs which is not enriched with repetitive DNA (Ibiapino et al., 2022). Thus, the accumulation of repetitive DNA can lead to an increase in chromosomes and, consequently, to an increase in genome size, especially in subgenus *Monogynella*.

Although genome size has been estimated for just a few species, the relative chromosome size may be an indirect indicator of genome size if we compare species with the same ploidy level. The smallest genome size in the genus estimated for the diploid *C. australis* ($2n = 30$; 1C = 0.27 Gbp) correlates well with the small chromosome size observed in mitotic metaphases (**Figures 1A,B**). Other diploids with the same chromosome number and similar chromosome size probably have a similar genome size. Such is the case of *C. desmouliniana* Yunck. (García et al., 2019) for which no genome size has been measured, but the size of chromosomes is similar or even smaller than in *C. australis*.

In the phylogeny of *Cuscuta*, some clades include species with noticeable differences in chromosome size. In subgenus *Pachystigma,* the number of large chromosomes varies from two pairs in *C. nitida* to five pairs in *C. angulata* suggesting a great difference in genome size within the subgenus. Further analysis of the two unsampled species from Eastern South Africa might reveal an even greater variation. Subgenus *Grammica*, accounting for ca. 70% of the *Cuscuta* species diversity (Stefanović et al., 2007), shows several clades with significant differences in chromosome and genome size between closely related species. A remarkable example is the strongly supported clade encompassing sections *Umbellatae*, *Indecorae*, and *Gracillimae* (clades L, M, N; García et al., 2014; Costea et al., 2015a), which has species with very small (*C. desmouliniana*, clade L), intermediate (*C. sidarum* and *C. globosa*, clade M), and large chromosomes (*C. coryli* Engelm. and *C. indecora*, clade M). The branch leading to species in sect. *Indecorae* is noticeably longer compared to others in the subgenus, showing an increase in mutation rates together with the increase of the genome size.

Although not as evident as in subgenus *Pachystigma*, some sections of subgenus *Grammica* (Costea et al., 2015a) appear to have a prevalence of asymmetrical karyotypes suggesting differential accumulation of DNA is some chromosome pairs (García et al., 2019) and probably to an increase in genome size. In section *Ceratophorae* (clade K; Costea et al., 2011), from which only diploid species are known, *C. costaricensis* Yunck. has one pair of chromosomes noticeably longer than the others, whereas in *C. bonafortunae* Costea & I. García or *C. erosa* Yunck. about half are long and half are shorter. Similar karyotypes have been documented in allopolyploids such as *C. veatchii*, but in these cases, the asymmetry is a consequence of the subgenomes of the two diploid parents with differences in chromosome size (Ibiapino et al., 2019).

Whereas some sections of subgenus *Grammica* show a tendency to increase genome size, others have undergone a significant reduction. The expansion and diversification of the genus in North America resulted in two lineages (Stefanović et al., 2007; García et al., 2014) having opposing directions in the evolution of genome size. The lineage of sections *Oxycarpae* (clade D) and *Denticulatae* (clade E) shows genome upsizing, which is especially evident in the former, including species with the biggest genomes in subgenus *Grammica* except for sect. *Indecorae*. On the contrary, the lineage that includes sections *Californicae* (clade A) and *Cleistogrammica* (clade B) has the smallest genomes not only in *Cuscuta* but also in Convolvulaceae. The only species of section *Racemosae* (clade C) with known karyotype and genome is *C. racemosa* ($2n = 60$; 1C = 1.39), having a relatively small genome compared to other tetraploids. Further sampling in this section will reveal whether the dispersal to South America before the diversification of this clade was accompanied by a significant variation in genome size.

Table 1 also summarizes the great intraspecific variation in the genome size of species of sections *Oxycarpae* and *Indecorae* (Costea et al., 2015a). Within *Oxycarpae*, differences such as in *C. compacta* (1C = 3.24 and 7.67) suggest that there might be diploid and tetraploid populations. In *C. gronovii*, all the chromosome counts are $2n = 60$; however, the genome size reported for the species ranges from 1C = 2.14 and 1C = 6.75 Gbp. Taxonomy of the section in general and of *C. gronovii* in particular is difficult and these differences might be explained by species identification mistakes or might reflect the morphological diversity of the species, for which several varieties have been described (Yuncker, 1932; Costea et al., 2006a). The strong support for the monophyly of the section contrasts with the very short internal branch lengths (Stefanović et al., 2007) suggesting a recent and rapid diversification accompanied by the accumulation of repetitive DNA with rates that might be different even at population level. A similar case is *C. indecora* in section *Indecorae*, for which several varieties have been described in addition to other two species, *C. coryli* and *C. warneri* Yunck. (Yuncker, 1932; Costea et al., 2006b).

The Increase of the Genome Size in Some Lineages of *Cuscuta* Is Probably Linked to Parasitism

Neumann et al. (2020) suggested that there is no correlation between genome size and the parasitic lifestyle of *Cuscuta*. However, according to the Plant DNA C-value Database, other species of Convolvulaceae have small genome sizes, with an average of approximately 1C = 0.97 Gbp and the largest value reported for the hexaploid *Ipomoea batatas* (L.) Lam. ($2n = 90$; 1C = 2.25 Gbp). Our estimation of the ancestral genome size of *Cuscuta* was ca. 1C = 12 Gbp, and 1C = 20 Gbp when including only *Cuscuta* species in the analyses. The phylogenetic position of subgenus *Monogynella* as sister to the rest of the genus, with such large genome sizes, may have resulted in an overestimation of the ancestral genome size for the genus as a whole. In spite of this possible bias, based on the limited Convolvulaceae data available, there was probably a genome expansion in subgenus *Monogynella*, and independent increases in other lineages, especially in sections *Indecorae* and *Oxycarpae* of subgenus *Grammica*.

The genome constraint hypothesis (Knight et al., 2005) suggests that the costs associated with the accumulation and replication of repetitive DNA reduce plant performance and negatively affects speciation and the distribution and abundance of species. Parasitic lifestyle eliminates the restrictions imposed by the growth rate of the meristem or the "genomic economy," because they take resources from their hosts (Gruner et al., 2010; Piednoël et al., 2012). Thus, despite the genomic reductions associated with the loss of autotrophic functions (Revill et al., 2005; Banerjee and Stefanović, 2019), there is a tendency for parasitic plants to have larger and more complex genomes (reviewed by Lyko and Wicke, 2021). For example, in Orobanchaceae, the genomes of the autotrophic *Lindenbergia philippensis* (Cham. & Schltdl.) Benth. and the hemiparasite *Schwalbea americana* L. are much smaller than those of the holoparasite *Orobanche* L. and *Phelipanche* Pomel, which fits the hypothesis of larger genome sizes in parasitic plants (Gruner et al., 2010; Piednoël et al., 2012). Without the selective constraints imposed by the nutrient and energy economy (Gruner et al., 2010; Piednoël et al., 2012), genome size could vary *via* mechanisms such as mobile elements activation and ectopic recombination. The amplification and diversity of transposable elements can be influenced by the DNA transposition and elimination rates, population size, reproduction mode, host plant defense mechanism, and even horizontal gene transfer (Devos et al., 2002; Bourque et al., 2018; Biscotti et al., 2019; Nishihara, 2020). This variation may or may not be fixed by the action of genetic drift and natural selection and could possibly allow a wider range of variation including the upper limits of genome size in *Cuscuta*, which are usually selected against in green plants.

Unlike in Orobanchaceae, genome size increases in *Cuscuta* are, however, not strictly related to the evolution to holoparasitism within the genus. Subgenus *Monogynella* has the least reduced plastome and higher ability for carbon fixation than the rest of subgenera; however, it has the largest genomes. Different levels of plastome reduction have been documented for the genus (Banerjee and Stefanović, 2020), and fully holoparasitic species are known in section *Ceratophorae* (subgenus *Grammica*, clade K; Banerjee and Stefanović, 2019). Although no genome size estimations have been done for the section, the chromosomes are not significantly bigger than in other related sections except for an apparent trend toward asymmetrical karyotypes. Species of sect. *Subulatae* (clade O) are also holoparasitic (Braukmann et al., 2013) and the only known genome sizes for the section are those of *C. chilensis* Ker Gawl. and *C. purpurata* Phil., two diploids with intermediate genome size (1C = 2.80 and 2.96, respectively).

Smaller genomes theoretically facilitate faster cell divisions and therefore growth (Gruner et al., 2010) and cell division rates (Símova and Herben, 2012). In *Cuscuta*, however, there is no clear negative correlation between genome size and growth rate, possibly because it may also be influenced by other factors, such as cell elongation. *Cuscuta indecora*, with the largest known genome in subgenus *Grammica*, is an invasive weed and a seed contaminant of crop plants, with profuse and fast growth over its hosts (Cudney et al., 1992; Costea et al., 2006b). Although we have not performed specific comparative experiments, we did not observe any evident difference in growth rate between this species and, e.g., *C. australis*, the species with the smallest genome known in the genus, even though both were growing on the same host in the same greenhouse conditions. It is remarkable that *C. indecora* behaves as a fast-growing weed, despite theoretically having longer cell cycles in which the whole genome must be replicated. *Cuscuta indecora* and weedy species of section *Oxycarpae,* such as *C. gronovii*, might be model systems to study the correlation between higher metabolic activity necessary to maintain the growth rate and the accumulation of repetitive DNA and transposable elements contributing to genome upsizing.

Changes in the Position of rDNA Sites May Indicate the Dynamics of Tandem Repetitive DNA Sequences in *Cuscuta*

Most *Cuscuta* species have a few rDNA sites: only one pair of 5S and one pair of 35S rDNA sites. Although the number of 5S and 35S rDNA loci is positively correlated with ploidy level (García et al., 2017), this does not hold true for this genus. *Cuscuta sandwichiana*, the highest polyploid reported, has one pair of 5S and two pairs of 35S rDNA sites, while phylogenetically close diploids, such as *C. australis*, have one pair of 5S and one pair of 35S rDNA sites. This may be due to the fact that some *Cuscuta* polyploids are interspecific hybrids, such as *C. sandwichiana* (Stefanović and Costea, 2008; García et al., 2014). The occurrence of recombination and gene conversion that results in the presence of rDNA copies from only one of the parents is often observed in hybrids. In allotetraploids of the *Dilatata* group of the genus *Paspalum* L. (Poaceae), for instance, the recovered ITS sequences show homogenization toward the paternal genome only (Vaio et al., 2019). In addition, the decrease in the number of expected sites may occur due to the elimination of some sites in terminal regions. The terminal position of the rDNA sites would be selectively favorable compared to the proximal ones, as it would reduce the chances of deleterious chromosomal rearrangements related to unequal recombination and recombination between non-homologous chromosomes (Roa and Guerra, 2012; García et al., 2017). Besides, the number of parental rDNA sites may be quite conserved in young, artificial allopolyploids. However, in natural allopolyploids, this number is often reduced, especially the 5S rDNA sites (Lee et al., 2011; Volkov, 2017). In *C. veatchii*, for example, there is a reduction in the rDNA sites in relation to its parents *C. denticulata* and *C. nevadensis* I.M. Johnst., indicating an old origin of this hybrid (Ibiapino et al., 2019).

It is common for the 5S and 35S rDNA sites to be found at separate locations in the genome, even on different chromosomes. This may be related to the fact that they are transcribed in different cellular compartments, by different enzymes (García and Kovařík, 2013). However, in *Cuscuta*, many species had rDNA sites located on the same chromosome, across all subgenera. In *C. monogyna* (subgenus *Monogynella*, sister to the rest of the genus), almost all its 30 chromosomes had 5S and 35S rDNA sites positioned closely, both on the same chromosome arm and on different arms. Co-occurrence of 5S and 35S rDNA sites on the same chromosome is higher in karyotypes with multiple sites and is frequently on the same arm (Roa and Guerra, 2015; García et al., 2017). However, all other *Cuscuta* species (e.g., *C. veatchii* and *C. indecora*) with sites on the same chromosome have these sites positioned on the same arm.

In plants, 5S rDNA usually occupies proximal and less frequently interstitial and terminal regions, while 35S rDNA tends to occupy terminal regions (Roa and Guerra, 2012, 2015). In *Cuscuta*, while 5S rDNA was more frequently found in interstitial regions, 35S is frequently found on peri/centromeric regions. In the only two holocentric species of *Cuscuta* whose rDNA sites are reported in this work, the positions of these sites also diverged from that found in other groups of plants. Generally, both 5S rDNA and 35S rDNA occupy terminal regions in holocentrics (Roa and Guerra, 2012, 2015). In holocentric *Cuscuta* species, only the 35S occupied a terminal position. The 5S rDNA sites were at more interstitial positions.

The ancestral character reconstruction suggested the interstitial position as ancestral for the 5S rDNA. This characteristic is present in all species, including *C. monogyna* (subgenus *Monogynella*) and *C. indecora* (subgenus *Grammica*), which also showed proximal and terminal sites, respectively. The 35S was more variable, and the "mix" condition, where rDNA sites were found in more than one location, was present at several clades throughout the phylogeny. The 35S rDNA is commonly pericentromeric; however, in species with more sites, additional sites are usually found in interstitial regions. In the subgenus *Cuscuta*, without a localized centromere, all 35S sites were terminal.

According to the Plant rDNA Database (García et al., 2017), within the Convolvulaceae, published rDNA site data are available only for seven species of *Ipomoea*, which is not closely related to *Cuscuta* and, thus, may not aid in resolving ancestral rDNA state. In the subgenus *Monogynella*, there is an increase in the diversity of positions in which the rDNA sites were found, suggesting that the "mix" condition is derived and, due to the amplification of these sites, the ribosomal DNAs began to occupy different positions along the chromosome. Recent studies show the possible influence of repetitive DNA amplification on genomic changes in the genus *Cuscuta* (Neumann et al., 2020; Ibiapino et al., 2022), indicating that the increase of these sites in *Monogynella* may be caused by the amplification of rDNA repeats as observed for other tandem repetitive sequences in the genus, and these are actually pseudogenes.

The tandem repetitive DNA in *Cuscuta* appears to be quite complex. For example, *C. europaea* have species-specific satellite DNA sequences such as the CUS-TR24, colocalized with centromeric proteins, also species-specific (Oliveira et al., 2020). Analysis using long reads showed a complex organization of CUS-TR24. The sequence of this satellite is interspersed with insertions mainly from LINE retrotransposons (Vondrak et al., 2021). In the subgenus *Pachystigma*, there is also evidence of these complex satellites. For example, the CnSat10-1,400 found in *C. nitida* in addition to being similar to a LINE type element, co-localizes with 35S signals in the chromosomes of this species. In addition, the most abundant SF1 family of *C. nitida* also co-locates with 35S signals (Ibiapino et al., 2022). This complex organization, with possible insertions of rDNA in other repetitive DNAs, could influence the diversity of number and position of rDNA sites in *Cuscuta*. In *Allium cepa* L., for example, it was shown that 35S rDNA is able to move from one locus to another in the genome. Furthermore, it is associated with telomeric DNA and other satellite DNAs, suggesting that the 35S of this species undergoes excision-reintegration mediated by these sequences (Mancia et al., 2015; Fu et al., 2019). Thus, the current data suggest that the evolution of rDNA in *Cuscuta*

may be influenced by other tandem repeats or transposable elements.

CONCLUSION

The data support the basic number $x = 15$ in *Cuscuta*, with duplications more common in subgenus *Grammica*. As expected, dysploidy occurred predominantly in the holocentric clade (subgenus *Cuscuta*). The remarkable increase and variation of the genome size in most lineages of *Cuscuta* may have been favored by the release of constraints enabled by its parasitic lifestyle. The data showed an expansion of genome size in comparison with the other Convolvulaceae, mostly by repetitive DNA amplification. This amplification of sequences may also have given rise to the great diversity of 5S and 35S ribosomal DNA sites found in the genus, and it seems to contribute to the emergence of "mix"-type karyotypes, in which multiple positions are occupied by these rDNA sites. This work analyzed data from 57 of the 200 *Cuscuta* species, which represents only 29% of *Cuscuta* species, indicating that the karyotypic diversity of the genus may be still greater than reported. Nevertheless, *Cuscuta* is one of the exceptionally diverse genera within the angiosperms in terms of karyotype and genome size. *Cuscuta*, having closely related species with different ploidy levels and marked differences in chromosome and genome size, is an excellent, tractable model system in which to study genome downsizing in polyploids as well as correlation of DNA content to phenotype such as pollen size, growth rate, cell cycle time, and epidermal cell size.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

AI performed the scientific experiments, data collection, and writing of the manuscript. MG contributed to the writing and reviewing of the manuscript and general discussions. BA performed all the phylogenetic analysis and supported the computational analysis. MB contributed by analyzing the data and reviewing the manuscript. MC and SS edited the manuscript, contributed to the collection, identification, and molecular data of the plant material, and discussions of the data. AP-H designed the experiments, supervised, and coordinated the project. All authors contributed to the article and approved the submitted version.

FUNDING

We thank the Fundação de Amparo a Ciencia e Tecnologia de Pernambuco (FACEPE) for the financing of the Postgraduate scholarship; the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq); and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Financial Code 001) for the financial support for the development of the project. BA thanks CAPES for the post-doc fellowship (process #88882.315044/2019-01). NSERC Discovery Canada supported the research of MC (327013) and SS (326439).

ACKNOWLEDGMENTS

We thank Marcelo Guerra (UFPE) for support at initial stages of this work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.842260/full#supplementary-material

REFERENCES

Ali, S. S., Yu, Y., Pfosser, M., and Wetschnig, W. (2012). Inferences of biogeographical histories within subfamily Hyacinthoideae using S-DIVA and Bayesian binary MCMC analysis implemented in RASP (reconstruct ancestral state in phylogenies). *Ann. Bot.* 109, 95–107. doi: 10.1093/aob/mcr274

Alix, K., Gérard, P. R., Schwarzacher, T., and Helsop-Harrison, J. S. (2017). Polyploidy and interspecific hybridization: partners for adaptation, speciation and evolution in plants. *Ann. Bot.* 120, 183–194. doi: 10.1093/aob/mcx079

Aryavand, A. (1987). The chromosome number of some Cuscuta (Cuscutaceae) species from Isfahan. *Iran. J. Bot.* 3, 177–182.

Bai, C., Alverson, W. S., Follansbee, A., and Waller, D. M. (2012). New reports of nuclear DNA content for 407 U.S. plant species. *Ann. Bot.* 110, 1623–1629. doi: 10.1093/aob/mcs222

Banerjee, A., and Stefanović, S. (2019). Caught in action: plastid genome evolution in *Cuscuta* sect. *Ceratophorae* (Convolvulaceae). *Plant Mol. Biol.* 100, 621–634. doi: 10.1007/s11103-019-00884-0

Banerjee, A., and Stefanović, S. (2020). Reconstructing plastome evolution across the phylogenetic backbone of the parasitic plant genus *Cuscuta* (Convolvulaceae). *Bot. J. Linn. Soc.* 194, 423–438. doi: 10.1093/botlinnean/boaa056

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., et al. (2012). GenBank. *Nucleic Acids Res.* 41, D36–D42. doi: 10.1093/nar/gks1195

Biscotti, M. A., Carducci, F., Olmo, E., and Canapa, A. (2019). "Vertebrate genome size and the impact of transposable elements in genome evolution," in *Evolution, Origin of Life, Concepts and Methods*. ed. P. Pontarotti (Cham: Springer)

Biswal, D. K., Debnath, M., Konhar, R., Yanthan, S., and Tandom, P. (2018). Phylogeny and biogeography of carnivorous plant family Nepenthaceae with reference to the Indian pitcher plant *Nepenthes khasiana* reveals an Indian subcontinent origin of *Nepenthes* colonization in South East Asia during the Miocene epoch. *Front. Ecol. Evol.* 6:108. doi: 10.3389/fevo.2018.00108

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., et al. (2018). Ten things you should know about transposable elements. *Genome Biol.* 19:199. doi: 10.1186/s13059-018-1577-z

Braukmann, T., Kuzmina, M., and Stefanović, S. (2013). Plastid genome evolution across the genus *Cuscuta* (Convolvulaceae): two clades within subgenus *Grammica* exhibit extensive gene loss. *J. Exp. Bot.* 64, 977–989. doi: 10.1093/jxb/ers391

Carta, A., Bedini, G., and Peruzzi, L. (2020). A deep dive into the ancestral chromosome number and genome size of flowering plants. *New Phytol.* 228, 1097–1106. doi: 10.1111/nph.16668

Costa, L., Oliveira, A., Carvalho-Sobrinho, J., and Souza, G. (2017). Comparative cytomolecular analyses reveal karyotype variability related to biogeographic and species richness patterns in Bombacoideae (Malvaceae). *Plant Syst. Evol.* 303, 1131–1144. doi: 10.1007/s00606-017-1427-6

Costea, M., García, M. A., and Stefanović, S. (2015a). A phylogenetically based infrageneric classification of the parasitic plant genus *Cuscuta* (Dodders, Convolvulaceae). *Syst. Bot.* 40, 269–285. doi: 10.1600/036364415X686567

Costea, M., García, M. A., Baute, K., and Stefanović, S. (2015b). Entangled evolutionary history of *Cuscuta pentagona* clade: a story involving hybridization and Darwin in the Galapagos. *Taxon* 64, 1225–1242. doi: 10.12705/646.7

Costea, M., Nesom, G. L., and Stefanović, S. (2006a). Taxonomy of *Cuscuta gronovii* and *Cuscuta umbrosa* (Convolvulaceae). *Sida* 22, 197–207.

Costea, M., Nesom, G. L., and Stefanović, S. (2006b). Taxonomy of the *Cuscuta indecora* (Convolvulaceae) complex in North America. *Sida* 22, 209–225.

Costea, M., Ruiz, I. G., and Stefanović, S. (2011). Systematics of "horned" dodders: phylogenetic relationships, taxonomy, and two new species within the *Cuscuta chapalana* complex (Convolvulaceae). *Botany* 89, 715–730. doi: 10.1139/b11-049

Costea, M., and Stefanović, S. (2010). Evolutionary history and taxonomy of the *Cuscuta umbellata* complex (Convolvulaceae): evidence of extensive hybridization from discordant nuclear and plastid phylogenies. *Taxon* 59, 1783–1800. doi: 10.1002/tax.596011

Cudney, D. W., Orloff, S. B., and Reints, J. S. (1992). An integrated weed management procedure for the control of dodder (*Cuscuta indecora*) in alfalfa (*Medicago sativa*). *Weed Technol.* 6, 603–606. doi: 10.1017/S0890037X00035879

Cusimano, N., Stadler, T., and Renner, S. S. (2012). A new method for handling missing species in diversification analysis applicable to randomly or nonrandomly sampled phylogenies. *Syst. Biol.* 61, 785–792. doi: 10.1093/sysbio/sys031

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

Devos, K. M., Brown, J. K., and Bennetzen, J. L. (2002). Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Res.* 12, 1075–1079. doi: 10.1101/gr.132102

Doležel, J., Greilhuber, J., and Suda, J. (2007). Estimation of nuclear DNA content in plants using flow cytometry. *Nat. Protoc.* 2, 2233–2244. doi: 10.1038/nprot.2007.310

Fogelberg, S. O. (1938). The cytology of *Cuscuta*. *Bull. Torrey. Bot. Club* 65:631. doi: 10.2307/2481064

Fu, J., Zhang, H., Guo, F., Ma, L., Wu, J., Yue, M., et al. (2019). Identification and characterization of abundant repetitive sequences in *Allium cepa*. *Sci. Rep.* 9:16756. doi: 10.1038/s41598-019-52995-9

García, M. A. (2001). A new western Mediterranean species of *Cuscuta* (Convolvulaceae) confirms the presence of holocentric chromosomes in subgenus *Cuscuta*. *Bot. J. Linn. Soc.* 135, 169–178. doi: 10.1111/j.1095-8339.2001.tb01089.x

García, M. A., and Castroviejo, S. (2003). Estudios citotaxonómicos en las especies Ibéricas del género *Cuscuta* (Convolvulaceae). *An. Jard. Bot. Madr.* 60, 33–44.

García, M. A., Costea, M., Guerra, M., García-Ruiz, I., and Stefanović, S. (2019). IAPT chromosome data 31. *Taxon* 68, 1374–1380. doi: 10.1002/tax.12176

García, M. A., Costea, M., Kuzmina, M., and Stefanović, S. (2014). Phylogeny, character evolution, and biogeography of *Cuscuta* (dodders; Convolvulaceae) inferred from coding plastid and nuclear sequences. *Am. J. Bot.* 101, 670–690. doi: 10.3732/ajb.1300449

García, S., and Kovařík, A. (2013). Dancing together and separate again: gymnosperms exhibit frequent changes of fundamental 5S and 35S rRNA gene (rDNA) organisation. *Heredity* 111, 23–33. doi: 10.1038/hdy.2013.11

García, S., Kovařík, A., Leitch, A. R., and Garnatje, T. (2017). Cytogenetic features of rRNA genes across land plants: analysis of the plant rDNA database. *Plant J.* 89, 1020–1030. doi: 10.1111/tpj.13442

García, M. A., and Martín, M. P. (2007). Phylogeny of *Cuscuta* subgenus *Cuscuta* (Convolvulaceae) based on nrDNA ITS and chloroplast *trnL* intron sequences. *Syst. Bot.* 32, 899–916. doi: 10.1600/036364407783390872

García, M. A., Stefanović, S., Weiner, C., Olszewski, M., and Costea, M. (2018). Cladogenesis and reticulation in *Cuscuta* sect. *Denticulatae* (Convolvulaceae). *Org. Divers. Evol.* 18, 383–398. doi: 10.1007/s13127-018-0383-5

Gerlach, W. L., and Bedbrook, J. R. (1979). Cloning and characterization of ribosomal RNA genes from wheat and barley. *Nucleic Acids Res.* 7, 1869–1885. doi: 10.1093/nar/7.7.1869

Glick, L., and Mayrose, I. (2014). ChromEvol: assessing the pattern of chromosome number evolution and the inference of polyploidy along a phylogeny. *Mol. Biol. Evol.* 31, 1914–1922. doi: 10.1093/molbev/msu122

Gruner, A., Hoverter, N., Smith, T., and Knight, C. A. (2010). Genome size is a strong predictor of root meristem growth rate. *J. Bot.* 2010, 1–4. doi: 10.1155/2010/390414

Guerra, M. (2008). Chromosome numbers in plant cytotaxonomy: concepts and implications. *Cytogenet. Genome Res.* 120, 339–350. doi: 10.1159/000121083

Guerra, M., and García, M. A. (2004). Heterochromatin and rDNA sites distribution in the holocentric chromosomes of *Cuscuta approximata* Bab. (Convolvulaceae). *Genome* 47, 134–140. doi: 10.1139/g03-098

Guo, S. L., Yu, J., Li, D. D., Zhou, P., Fang, Q., and Yin, L. P. (2015). DNA C-values of 138 herbaceous species and their biological significance. *Acta Ecol. Sin.* 35, 6516–6529. doi: 10.5846/stxb201406111208

Heslop-Harrison, J. S. P., and Schwarzacher, T. (2011). Organisation of the plant genome in chromosomes. *Plant J.* 66, 18–33. doi: 10.1111/j.1365-313X.2011.04544.x

Ho, A., and Costea, M. (2018). Diversity, evolution and taxonomic significance of fruit in *Cuscuta* (dodder, Convolvulaceae); the evolutionary advantages of indehiscence. *Perspect. Plant Ecol.* 32, 1–17. doi: 10.1016/j.ppees.2018.02.001

Ibiapino, A., Baez, M., García, M. A., Costea, M., Stefanovic, S., and Pedrosa-Harand, A. (2022). Karyotype asymmetry in *Cuscuta* L. subgenus *Pachystigma* reflects its repeat DNA composition. *Chrom. Res.* doi: 10.1007/s10577-021-09683-0

Ibiapino, A., García, M. A., Costea, M., Stefanović, S., and Guerra, M. (2020). Intense proliferation of rDNA sites and heterochromatic bands in two distantly related *Cuscuta* species (Convolvulaceae) with very large genomes and symmetric karyotypes. *Genet. Mol. Biol.* 43:e20190068. doi: 10.1590/1678-4685-gmb-2019-0068

Ibiapino, A., García, M. A., Ferraz, M. E., Costea, M., Stefanović, S., and Guerra, M. (2019). Allopolyploid origin and genome differentiation of the parasitic species *Cuscuta veatchii* (Convolvulaceae) revealed by genomic in situ hybridization. *Genome* 62, 467–475. doi: 10.1139/gen-2018-0184

Ishikawa, S. A., Zhukova, A., Iwasaki, W., and Gascuel, O. (2019). A fast likelihood method to reconstruct and visualize ancestral scenarios. *Mol. Biol. Evol.* 36, 2069–2085. doi: 10.1093/molbev/msz131

Kaul, M. L. H., and Bhan, A. K. (1977). Cytogenetics of polyploids: VI. Cytology of tetraploid and hexaploid *Cuscuta reflexa* Roxb. *Cytologia* 42, 125–136. doi: 10.1508/cytologia.42.125

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

Khatoon, S., and Ali, S. I. (1993). *Chromosome Atlas of the Angiosperms of Pakistan*. Karachi: Department of Botany, University of Karachi.

Knight, C. A., Molinari, N. A., and Petrov, D. A. (2005). The large genome constraint hypothesis: evolution, ecology and phenotype. *Ann. Bot.* 95, 177–190. doi: 10.1093/aob/mci011

Lee, Y.-I., Chang, F.-C., and Chung, M.-C. (2011). Chromosome pairing affinities in interspecific hybrids reflect phylogenetic distances among lady's slipper orchids (*Paphiopedilum*). *Ann. Bot.* 108, 113–121. doi: 10.1093/aob/mcr114

Leitch, A. R., and Leitch, I. J. (2008). Genomic plasticity and the diversity of polyploid plants. *Science* 320, 481–483. doi: 10.1126/science.1153585

Loureiro, J., Rodriguez, E., Dolezel, J., and Santos, C. (2007). Two new nuclear isolation buffers for plant DNA flow cytometry: a test with 37 species. *Ann. Bot.* 100, 875–888. doi: 10.1093/aob/mcm152

Love, D. (1982). IOPB Chromosome Number reports LXXVI. *TAXON*. 31, 583–587.

Lyko, P., and Wicke, S. (2021). Genomic reconfiguration in parasitic plants involves considerable gene losses alongside global genome size inflation and gene births. *Plant Physiol.* 186, 1412–1423. doi: 10.1093/plphys/kiab192

Maddison, W. P., and Maddison, D. R. (2018). Mesquite: a modular system for evolutionary analysis. Version 2.75. Available at: http://www.mesquiteproject.org (Accessed February 24 2022).

Mancia, F. H., Sohn, S.-H., Ahn, Y. K., Kim, D.-S., Kim, J. S., Kwon, Y.-S., et al. (2015). Distribution of various types of repetitive DNAs in *Allium cepa* L. based on dual color FISH. *Hortic. Environ. Biotechnol.* 56, 793–799. doi: 10.1007/s13580-015-1100-3

Mandáková, T., and Lysak, M. A. (2018). Post-polyploid diploidization and diversification through dysploid changes. *Curr. Opin. Plant Biol.* 42, 55–65. doi: 10.1016/j.pbi.2018.03.001

Mandrioli, M., and Manicardi, G. C. (2020). Holocentric chromosomes. *PLoS Genet.* 16:e1008918. doi: 10.1371/journal.pgen.1008918

Márquez-Corro, J. I., Martín-Bravo, S., Spalink, D., Luceño, M., and Escudero, M. (2019). Inferring hypothesis-based transitions in clade-specific models of chromosome number evolution in sedges (Cyperaceae). *Mol. Phylogenet. Evol.* 135, 203–209. doi: 10.1016/j.ympev.2019.03.006

Mayrose, I., and Lysak, M. A. (2021). The evolution of chromosome numbers: mechanistic models and experimental approaches. *Genome Biol. Evol.* 13:evaa220. doi: 10.1093/gbe/evaa220

McNeal, J. R., Kuehl, J. V., Boore, J. L., and de Pamphilis, C. W. (2007). Complete plastid genome sequences suggest strong selection for retention of photosynthetic genes in the parasitic plant genus *Cuscuta*. *BMC Plant Biol.* 7:57. doi: 10.1186/1471-2229-7-57

Mehra, P. N., and Vasudevan, K. N. (1972). IOPB Chromosome Number reports, XXXVI. *TAXON*. 21, 341–344.

Miller, M. A., Pfeiffer, W., and Schwartz, T. (2010). "Creating the CIPRES science gateway for inference of large phylogenetic trees." in *Gateway Computing Environments Worshop (GCE)*, 1–8. Available at: http://www.phylo.org/index.php/ (Accessed October 15 2022).

Mukerjee, S. K., and Bhattacharya, P. K. (1970). A new *Cuscuta* from Bengal. *Bull. Bot. Soc. Bengal* 24, 147–149.

Neumann, P., Oliveira, L., Čížková, J., Jang, T.-S., Klemme, S., Novák, P., et al. (2020). Impact of parasitic lifestyle and different types of centromere organization on chromosome and genome evolution in the plant genus *Cuscuta*. *New Phytol.* 229, 2365–2377. doi: 10.1111/nph.17003

Nishihara, H. (2020). Transposable elements as genetic accelerators of evolution: contribution to genome size, gene regulatory network rewiring and morphological innovation. *Genes Genet. Syst.* 94, 269–281. doi: 10.1266/ggs.19-00029

Oliveira, L., Neumann, P., Jang, T.-S., Klemme, S., Schubert, V., Koblížková, A., et al. (2020). Mitotic spindle attachment to the holocentric chromosomes of *Cuscuta europaea* does not correlate with the distribution of CENH3 chromatin. *Front. Plant Sci.* 10:1799. doi: 10.3389/fpls.2019.01799

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

Pazy, B., and Plitmann, U. (1994). Holocentric chromosome behaviour in *Cuscuta* (Cuscutaceae). *Plant Syst. Evol.* 191, 105–109. doi: 10.1007/BF00985345

Pazy, B., and Plitmann, U. (1995). Chromosome divergence in the genus *Cuscuta* and its systematic implications. *Caryologia* 48, 173–180. doi: 10.1080/00087114.1995.10797327

Pazy, B., and Plitmann, U. (2002). New perspectives on the mechanisms of chromosome evolution in parasitic flowering plants. *Bot. J. Linn. Soc.* 138, 117–122. doi: 10.1046/j.1095-8339.2002.00006.x

Pedrosa, A., Sandal, N., Stougaard, J., Schweizer, D., and Bachmair, A. (2002). Chromosomal map of the model legume *Lotus japonicus*. *Genetics* 161, 1661–1672. doi: 10.1093/genetics/161.4.1661

Piednoël, M., Aberer, A. J., Schneeweiss, G. M., Macas, J., Novak, P., Gundlach, H., et al. (2012). Next-generation sequencing reveals the impact of repetitive DNA across phylogenetically closely related genomes of Orobanchaceae. *Mol. Biol. Evol.* 29, 3601–3611. doi: 10.1093/molbev/mss168

Pustahija, F., Brown, S. C., Bogunic, F., Bašic, N., Muratovic, E., Ollier, S., et al. (2013). Small genomes dominate in plants growing on serpentine soils in West Balkans, an exhaustive study of 8 habitats covering 308 taxa. *Plant Soil* 373, 427–453. doi: 10.1007/s11104-013-1794-x

Qiu, T., Liu, Z., and Liu, B. (2020). The effects of hybridization and genome doubling in plant evolution via allopolyploidy. *Mol. Biol. Rep.* 47, 5549–5558. doi: 10.1007/s11033-020-05597-y

R Core Team (2020). R: a language and environment for statistical computing. Available at: http://www.Rproject.org/ (Accessed February 09 2022).

Rambaut, A. (2014). FigTree v1.4.2. Available at: http://tree.bio.ed.ac.uk/software/Figtree/ (Accessed February 23 2022).

Rambaut, A., Suchard, M. A., Xie, W., and Drummond, A. J. (2014). TRACER v.1.6, Available at: http://tree.bio.ed.ac.uk/software/tracer/ (Accessed February 09 2022).

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

Revill, M. J. W., Stanley, S., and Hibberd, J. M. (2005). Plastid genome structure and loss of photosynthetic ability in the parasitic genus *Cuscuta*. *J. Exp. Bot.* 56, 2477–2486. doi: 10.1093/jxb/eri240

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

Rice, A., and Mayrose, I. (2021). Model adequacy tests for probabilistic models of chromosome number evolution. *New Phytol.* 229, 3602–3613. doi: 10.1111/nph.17106

Roa, F., and Guerra, M. (2012). Distribution of 45S rDNA sites in chromosomes of plants: structural and evolutionary implications. *BMC Evol. Biol.* 12:225. doi: 10.1186/1471-2148-12-225

Roa, F., and Guerra, M. (2015). Non-random distribution of 5S rDNA sites and its association with 45S rDNA in plant chromosomes. *Cytogenet. Genome Res.* 146, 243–249. doi: 10.1159/000440930

Roberto, C. (2005). Low chromosome number angiosperms. *Caryologia* 58, 403–409. doi: 10.1080/00087114.2005.10589480

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

Sader, M. A., Amorim, B. S., Costa, L., Souza, G., and Pedrosa-Harand, A. (2019). The role of chromosome changes in the diversification of *Passiflora* L. (Passifloraceae). *Syst. Biodivers.* 17, 7–21. doi: 10.1080/14772000.2018.1546777

Schubert, I., and Lysak, M. A. (2011). Interpretation of karyotype evolution should consider chromosome structural constraints. *Trends Genet.* 27, 207–216. doi: 10.1016/j.tig.2011.03.004

Símova, I., and Herben, T. (2012). Geometrical constraints in the scaling relationships between genome size, cell size and cell cycle length in herbaceous plants. *Proc. Biol. Sci.* 279, 867–875. doi: 10.1098/rspb.2011.1284

Stefanović, S., and Costea, M. (2008). Reticulate evolution in the parasitic genus *Cuscuta* (Convolvulaceae): over and over. *Botany* 86, 791–808. doi: 10.1139/B08-033

Stefanović, S., Kuzmina, M., and Costea, M. (2007). Delimitation of major lineages within *Cuscuta* subgenus *Grammica* (dodders; Convolvulaceae) using plastid and nuclear DNA sequences. *Am. J. Bot.* 94, 568–589. doi: 10.3732/ajb.94.4.568

Stefanović, S., and Olmstead, R. G. (2004). Testing the phylogenetic position of a parasitic plant (*Cuscuta*, Convolvulaceae, Asteridae): Bayesian inference and the parametric bootstrap on data drawn from three genomes. *Syst. Biol.* 53, 384–399. doi: 10.1080/10635150490445896

Suda, J., Kyncl, T., and Freiova, R. (2003). Nuclear DNA amounts in Macaronesian angiosperms. *Ann. Bot.* 92, 153–164. doi: 10.1093/aob/mcg104

Suda, J., Kyncl, T., and Jarolimova, V. (2005). Genome size variation in Macaronesian angiosperms: forty percent of the Canarian endemic flora completed. *Plant Syst. Evol.* 252, 215–238. doi: 10.1007/s00606-004-0280-6

Sun, G., Xu, Y., Liu, H., Sun, T., Zhang, J., Hettenhausen, C., et al. (2018). Large-scale gene losses underlie the genome evolution of parasitic plant *Cuscuta australis*. *Nat. Commun.* 9:2683. doi: 10.1038/s41467-018-04721-8

Uhl, C. H. (1978). Chromosomes of Mexican *sedum*. II section pachysedum. *Rhodora* 80, 491–512.

Vaio, M., Gardner, A., Emshwiller, E., and Guerra, M. (2013). Molecular phylogeny and chromosome evolution among the creeping herbaceous

Oxalis species of sections *Corniculatae* and *Ripariae* (Oxalidaceae). *Mol. Phylogenet. Evol.* 68, 199–211. doi: 10.1016/j.ympev.2013.03.019

Vaio, M., Mazzella, C., Guerra, M., and Speranza, P. (2019). Effects of the diploidisation process upon the 5S and 35S rDNA sequences in the allopolyploid species of the Dilatata group of *Paspalum* (Poaceae, Paniceae). *Aust. J. Bot.* 67:521. doi: 10.1071/BT18236

Veleba, A., Šmarda, P., Zedek, F., Horová, L., Šmerda, J., and Bureš, P. (2017). Evolution of genome size and genomic GC content in carnivorous holokinetics (Droseraceae). *Ann. Bot.* 119, 409–416. doi: 10.1093/aob/mcw229

Veleba, A., Zedek, F., Horová, L., Veselý, P., Srba, M., Šmarda, P., et al. (2020). Is the evolution of carnivory connected with genome size reduction? *Am. J. Bot.* 107, 1253–1259. doi: 10.1002/ajb2.1526

Vogel, A., Schwacke, R., Denton, A. K., Usadel, B., Hollmann, J., Fischer, K., et al. (2018). Footprints of parasitism in the genome of the parasitic flowering plant *Cuscuta campestris*. *Nature Comm.* 9:2515 doi: 10.1038/s41467-018-04344-z

Volkov, R. A. (2017). Evolutional dynamics of 45S and 5S ribosomal DNA in ancient allohexaploid *Atropa belladonna*. *BMC Plant Biol.* 17:21. doi: 10.1186/s12870-017-0978-6

Vondrak, T., Oliveira, L., Novák, P., Koblížková, A., Neumann, P., and Macas, J. (2021). Complex sequence organization of heterochromatin in the holocentric plant *Cuscuta europaea* elucidated by the computational analysis of nanopore reads. *Comput. Struc. Biotech. J.* 19, 2179–2189. doi: 10.1016/j.csbj.2021.04.011

Weiss-Schneeweiss, H., Greilhuber, J., and Schneeweiss, G. M. (2006). Genome size evolution in holoparasitic *Orobanche* (Orobanchaceae) and related genera. *Am. J. Bot.* 93, 148–156. doi: 10.3732/ajb.93.1.148

Yoshida, K., and Kitano, J. (2021). Tempo and mode in karyotype evolution revealed by a probabilistic model incorporating both chromosome number and morphology. *PLoS Genet.* 17:e1009502. doi: 10.1371/journal.pgen.1009502

Yu, Y., Blair, C., and He, X. (2020). RASP 4: ancestral state reconstruction tool for multiple genes and characters. *Mol. Biol. Evol.* 37, 604–606. doi: 10.1093/molbev/msz257

Yu, Y., Harris, A. J., Blair, C., and He, X. (2015). RASP (reconstruct ancestral state in phylogenies): a tool for historical biogeography. *Mol. Phylogenet. Evol.* 87, 46–49. doi: 10.1016/j.ympev.2015.03.008

Yuncker, T. G. (1932). The genus *Cuscuta*. *Mem. Torrey Bot. Club* 18, 113–331.

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

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

Copyright © 2022 Ibiapino, García, Amorim, Baez, Costea, Stefanović and Pedrosa-Harand. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 07 April 2022
doi: 10.3389/fpls.2022.860778

Evaluation of Four Commonly Used DNA Barcoding Loci for *Ardisia* Species Identification

Chao Xiong[1†], Wei Sun[2†], Lan Wu[2], Ran Xu[1], Yancheng Zhang[3], Wenjun Zhu[1], H. E. J.[4], Panjwani[5], Zhiguo Liu[1] and Bo Zhao[2]**

[1]School of Life Science and Technology, Wuhan Polytechnic University, Wuhan, China, [2]Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing, China, [3]Department of Pharmacognosy, Pharmacy School, Guilin Medical University, Guilin, China, [4]Research Institute of Chemistry, International Center for Chemical and Biological Sciences, University of Karachi, Karachi, Pakistan, [5]Center for Molecular Medicine and Drug Research, International Center for Chemical and Biological Sciences, University of Karachi, Karachi, Pakistan

OPEN ACCESS

Edited by:
Isabel Mafra,
University of Porto, Portugal

Reviewed by:
Velusamy Sundaresan,
Council of Scientific and Industrial Research (CSIR), India
Zhi Chao,
Southern Medical University, China

***Correspondence:**
Zhiguo Liu
zhiguo_l@126.com
Bo Zhao
2052886016@qq.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: 25 January 2022
Accepted: 18 March 2022
Published: 07 April 2022

Citation:
Xiong C, Sun W, Wu L, Xu R, Zhang Y, Zhu W, J. HE, Panjwani, Liu Z and Zhao B (2022) Evaluation of Four Commonly Used DNA Barcoding Loci for Ardisia Species Identification. Front. Plant Sci. 13:860778. doi: 10.3389/fpls.2022.860778

Ardisia plants have been used as medicinal plants for a long time in China. Traditional techniques such as morphological, microscopic, and chemical identification methods all have limitations in the species identification of *Ardisia*. For the sake of drug safety, four DNA barcodes (*psbA-trnH*, ITS, *rbcL*, and *matK*) were assessed for Chinese *Ardisia* plants using a total of 121 individuals from 33 species. Four criteria (The success rates of PCR amplification, DNA barcoding gap, DNA sequence similarity analysis and NJ tree clustering analysis) were used to evaluate the species identification ability of these four DNA barcodes. The results show that ITS had the highest efficiency in terms of PCR and sequencing and exhibited the most apparent inter- and intra-specific divergences and the highest species identification efficiency. There was no significant increase in species identification after combining the three cpDNA fragments with the ITS fragment. Considering the cost and experimental effectiveness, we recommend ITS as the core barcode for identifying Chinese *Ardisia* plants.

Keywords: *Ardisia*, DNA barcoding, species identification, ITS fragment, cpDNA fragment

INTRODUCTION

Ardisia, which comprises approximately 500 species worldwide and 65 species, have been recorded in the latest publication of "Flora of China" (Chen and Pipoly, 1996). *Ardisia* species have been used as medicine, food and ornamental plants for a long time. Because of their high medicinal and aesthetic value, *Ardisia* species have a sizeable exploratory potential and a broad market foreground (Liu et al., 2013). The dried plants of several *Ardisia* species have been used as Chinese Traditional Medicine for the treatment of ailments such as conjunctivitis, bronchitis, pneumonia, tuberculosis trauma, as well as pancreatic and other types of cancer (Zhao et al., 2014; François et al., 2016; Oliveira et al., 2018; Sanjeev et al., 2019). For example, the herb Aidicha found in China, the whole dried plant of *Ardisia japonica* (Thunberg) Blume, is officially listed in the Chinese Pharmacopoeia.

Since the 1970s, extensive research on the chemical components and pharmacological action of *Ardisia* has resulted in the discovery of many novel biologically active ingredients.

Kobayashi and de Mejía (2005) have reviewed the chemical composition, biological activity, and pharmacological effects of many *Ardisia* plants. *Ardisia* species contain various physiologically active compounds, such as peptides, saponins, isocoumarins, quinones and alkylphenols I; therefore, it has high medicinal value with antitussive, anti-fertility, antiasthmatic, anti-inflammatory, antibacterial, anti-viral, anti-tumor and insecticidal effects (Newell et al., 2010; Joaquín-Cruz et al., 2015). Based on these pharmacological, now various products such as Aidicha Capsules, Compound Aidicha Tablets and Zouchuan Guci Ding have already been produced and used in clinical applications in China (Xin et al., 2015); in addition, extracts of *Ardisia colorata* Roxb are also commonly used in Thailand to treat gastrointestinal infections (Voravuthikunchai et al., 2004).

Despite a long history of its medicinal use in China, the taxonomic ambiguities make proper identification and acquisition of plant materials of some *Ardisia* species difficult. Mistaken identification is also due to either confusing nomenclature or several common terms with transliterated and local names, to the extent that some areas have 5 or 6 different names for the same species. Besides taxonomic confusion, the safety and quality of *Ardisia* herbal products have been a matter of increasing concern (Liu et al., 2013). Manufacturers may be tempted to label their food products incorrectly and add lower-priced ingredients of inferior quality to increase their profit because suspect or counterfeit herbal materials have been found on sale. Market surveys identified the low-cost herbs *Rhododendron molle* root and *Clerodendrum cyrtophyllum* stem as the primary adulterated materials in the commercial products of *A. gigantifolia* (Dai et al., 2018). Therefore, the safety and quality of *Ardisia* herbal products need to be urgently addressed to protect customer health and maintain the quality and authenticity of these herbal products in the drug supply chain.

While morphological, microscopic, and chemical identification methods are primarily used to authenticate herbal materials from *Ardisia*, all of these traditional techniques have limitations. Besides the morphological similarity and variation in sample profiles, the accuracy of these methods also lies in the assessor's expertise. In addition, convergent evolution and extensive intra-species morphological variation make it laborious to identify and classify *Ardisia* species. It is also challenging to identify the botanical origin when analyzing heavily processed plant material. Recent developments in molecular biology and molecular genetic techniques have enabled the identification and authentication of *Ardisia* species. DNA-based methods are widely used in different research fields because they are rapid and sensitive. DNA barcoding, developed by the Centre for Biodiversity Genomics at the University of Guelph (Canada), is a powerful tool for species identification (Hebert et al., 2003a,b). This method is not limited by physiological conditions and morphological characteristics of samples, allowing species identification even without specialized taxonomic knowledge. The method can also be standardized for specific DNA barcodes and universal primers, a characteristic that is advantageous for building databases and creating a universal standard for identification (Chen et al., 2010; Yao et al., 2010; Li et al., 2011; Poudel et al., 2011). This method can identify species rapidly and accurately from a broad range and variable quality of raw materials and has a huge application potential in the food and medicine industries, which mainly ensures the use of correct, uncontaminated, and unsubstituted herbal ingredients (Chen et al., 2014). DNA barcode has become one of the most important tools for medicinal plant taxonomy and is used to identify adulterants in commercial herbal products (Cui et al., 2020; Yang et al., 2020), could regulate the quality in raw herbal trade market (Santhosh Kumar et al., 2015; Skjua et al., 2020).

Variation within a standard region of the genome called "DNA barcode" is analyzed using the DNA barcoding approach. This short sequence, which is derived from a suitable segment of the mitochondrial, chloroplast, or nuclear genome, is used to identify organisms at the species level. In 2009, after analyzing seven plastid DNA regions, including *atp*F-*atp*H, *mat*K, *rbc*L, *rpo*B, *rpo*C1, *psb*A-*trn*H, and *psb*K-*psb*I in 907 samples of 550 species, a combination of chloroplast Maturase K (*mat*K) and ribulose-bisphosphate carboxylase (*rbc*L) has been recommended as the core barcode for land plants (CBOL Plant Working Group, 2009). Subsequently, the chloroplast *psb*A-*trn*H region and the internal transcribed spacer (ITS) of nuclear ribosomal DNA were also considered for the core barcode of seed plants (Fazekas et al., 2008; China Plant BOL Group, 2011; Amritha et al., 2020; Zhang and Jiang, 2020). Usually, in plants, while the *mat*K region and the intergenic spacer *psb*A-*trn*H have evolved rapidly, the evolution of the *rbc*L region has not been so swift. In addition, a barcoding locus, the ITS region was more conservative than them, and all of them have been effectively used for complex plant groups. Although single or combined loci were used for candidate barcode sequences for plant identification, the most suitable DNA barcode for specific groups must be chosen by sequencing and analysis (Hollingsworth et al., 2011).

In *Ardisia*, the development of DNA barcoding is still nascent with few studies examining DNA regions to discriminate between species. This study used four core DNA barcodes (ITS, *mat*K, *rbc*L and *psb*A-*trn*H) to identify Chinese *Ardisia* plants. In this study, we aimed to assess the utility of these regions in Chinese *Ardisia* species and identify and screen out the best sequence suitable for applying DNA barcode technology in Chinese *Ardisia* species discrimination.

MATERIALS AND METHODS

Taxon Sampling

Based on the field investigation, 121 samples from 33 wild Chinese *Ardisia* species were included in this study. For molecular analyses, fresh leaves were randomly collected in the squaring period and desiccated in silica gel. *Embelia laeta*, *E. rudis* and *Glaux maritima* were selected as outgroups. The *Ardisia* species used in this study are listed in **Supplementary Table 1**. We verified the identity of all of the samples independently through consultation with expert Shizhong Mao, who is an Associate Researcher at the Guangxi Institute of Botany, Chinese

Academy of Sciences. Voucher specimens have been deposited in the Guangxi Institute of Botany, Chinese Academy of Sciences.

DNA Isolation, Amplification and Sequencing

For each sample, 30–40 mg of leaves dried by silica gel were used, and genomic DNA was extracted and purified according to the Plant Genomic DNA Kit (Tiangen Biotech Co., China). The DNA concentration was estimated using BioTek Epoch (BioTek, Co., United States) by standard spectrophotometric methods at 260 and 280 nm. DNA integrity was assessed by electrophoresis using 1.0% agarose gel. Then, the DNA samples were diluted to a working concentration of 50 ng/μl and stored at −20°C until further use. According to Zhang et al. (2012), four commonly used DNA barcoding loci (*mat*K, *rbc*L, *psb*A-*trn*H, and ITS) were used in this study. The steps were carried out according to the China Plant BOL Group (2011) and Chen et al. (2010) using DNA barcoding standard operating procedures (DNA barcoding SOP).

PCR amplification was performed in 25 μl reaction mixtures containing 20–50 ng of genomic DNA, 12.5 μl of 2 × Taq PCR MasterMix (Beijing Aidlab Biotech Co., Beijing, China), 1 μl of 2.5 μM forward and reverse primers, and distilled water up to the final volume. PCR products were assessed on 1.0% agarose gel, visualized under UV light, purified using a Multifunction DNA Purification Kit from Bioteke (China) and then sequenced in both directions on a 3730XL sequencer (Applied Biosystems, United States) using amplification primers listed in **Supplementary Table 2** (Chen, 2015).

Sequence Analyses

The sequences were proofread, assembled as contigs and consensus sequences were generated using the CodonCode Aligner 4.2.1 (CodonCode Co., Dedham, MA, United States). Then, the Basic Local Alignment Search Tool (BLAST, NCBI) was used to check the homology of the obtained sequences. The CLUSTAL X 2.0 (Larkin et al., 2007) was applied using the default parameters and then manually rectified for multiple nucleotide alignment. The base compositions, the genetic distances, variable sites and parsimony-informative site values were estimated by MEGA5.1 as per the K2P (Kimura 2 parameter) model (Tamura et al., 2011). Barcoding gaps were estimated by comparing the distributions of intra- and inter-specific divergences of each candidate locus using the program MEGA5.1.

The degree of species resolution (identification) for the four DNA barcode regions was evaluated using the NJ tree method. For each sequence data set, pairwise genetic distances and all possible combinations for the five sequence data sets were determined by the K2P (Kimura 2-parameter) method (Kimura, 1980) using MEGA5.1. Support for clades was evaluated by bootstrap analysis with 1,000 replicates. Species discrimination was considered successful only when a single clade in NJ (Neighbour Joining) trees with a bootstrap value above 50% was specifically formed by all conspecific individuals (Zhang et al., 2012).

Based on the analysis of DNA sequence similarity results, the Taxon DNA method was used to assess each barcode region and their probable combinations to determine the degree of species resolution they presented (Meier et al., 2006). Additionally, the "best match" and the "best close match" functions and the Taxon DNA method were applied to test the individual-level discrimination rates for each single marker and all possible combinations under the K2P-corrected distance model. The "best match" was used to search the closest barcode match for each query. The identification was deemed successful if both sequences were from the same species, whereas mismatched names were failures. However, if there were several equally valid 'best matches' from different species, they were considered ambiguous. The "best close match" was used to plot the relative frequency of intraspecific distances, and the threshold value less than 95% of all intraspecific distances was set. Each query that did not have a barcode match below the threshold value could not be identified. For the remaining queries, their identities were compared with the species identities of their closest barcodes. If the name was identical, the query was considered successful identification. The query was considered a failure when the names were mismatched and ambiguous when several equally valid best matches belonged to a minimum of two species (Meier et al., 2006; Zhang et al., 2012).

Finally, the NCBI BLAST program 2.2.29 + (Tao, 2010) was used for all sequences analyzed using the "BLASTn" command to build local reference databases. Successful species discrimination was deemed when all species had the highest hit matching only a conspecific individual; for better clarity, the query sequence was removed from the list of top hits (Meyer and Paulay, 2005).

RESULTS

PCR and Sequence Analysis

Total genomic DNA was extracted from 121 samples representing 33 Chinese Ardisia species, and then PCR and sequencing were carried out. All 468 sequences, including those of 119 matk, 114 rbcL, 121 ITS, and 114 trnH-psbA sequences, were obtained in this study, and submitted to the GenBank database (**Supplementary Table 1**). The efficiency of PCR amplification, in descending order, was 100.00% (ITS), 98.35% (matK), 94.21% (rbcL), and 94.21% (psbA-trnH). The failed species for matK were *A. arborescens*. The failed species for rbcL and psbA-trnH were *A. arborescens*, *A. humilis,* and *A. obtusa*. The sequencing for all four loci had a 100.0% success rate (**Table 1**).

The summary of the sequence characteristics of the four regions is presented in **Table 1**. The ITS sequences were 599–611 bp long, with 54.9–58.8% GC content and its multiple sequence alignment consisted of 619 characters, while 158 of 229 variable sites were potentially informative of parsimony. The matK sequences were 848–856 bp long, with 32.4–34.5% GC content and its multiple sequence alignment consisted of 856 characters, while 39 of 87 variable sites were potentially informative of parsimony. The rbcL sequence was 705 bp long with 30.9% GC content and its multiple sequence alignment

TABLE 1 | Success rates for PCR amplification and sequencing, and sequence characteristics of each single candidate barcodes.

	ITS	*mat*K	*rbc*L	*psb*A-*trn*H
Number of samples (individuals)	121	119	114	114
Success rates for PCR amplification (%)	100	98.35	94.21	94.21
Success rates for sequencing (%)	100	100	100	100
Length range (bp)	599–611	848–856	705	378–454
Aligned sequence length (bp)	619	856	705	533
GC content (%)	54.9–58.8	32.4–34.5	43.0–43.7	25.6–28.0
No. variable sites	229	87	40	85
No. parsimony information variable sites	158	39	19	47
Mean inter-specific distance (range), %	0.09 (0–1.93)	0.03 (0–0.59)	0.08 (0–0.85)	0.11 (0–6.56)
Mean intra-specific distance (range), %	3.43 (0–7.59)	0.39 (0–1.59)	0.37 (0–1.41)	1.49 (0–7.56)

consisted of 705 characters, while 19 of 40 variable sites were potentially informative of parsimony. The psbA-trnH sequence was 378–454 bp long with 25.6–28.0% GC content and its multiple sequence alignment consisted of 533 characters, while 47 of 85 variable sites were potentially informative parsimony.

Barcoding Gap Test

In an ideal DNA barcode, the bar chart for divergence must show an intra-specific variation with the focus on the left side, having smaller numerals, while inter-specific variation should have the focus on the right side with greater numerals (Priyanka et al., 2015). This study observed obvious gaps between the intra- and inter-specific variability of ITS, although there was a slight overlap. For the other three loci, no evident barcoding gaps were found, and the overlap in the distributions of intra- and inter-specific variation was obvious (**Table 1** and **Figure 1**).

Species Discrimination Using NJ Tree Method and Local BLAST Method

The NJ tree method was used to assess the identification efficiency at the species level for Chinese Ardisia based on the four candidate barcodes (individually and in combination). If individuals of a species all formed a monophyletic clade, it was considered a successful identification at the species level. The results showed that species discrimination levels for the ITS markers and seven combinations containing ITS sequence were 79.3–83.8%, while species discrimination levels for the other three markers (matK, rbcL, and psbA-trnH) and four combinations without ITS sequence were less than 60%.

In the NJ trees based on ITS markers and the barcodes combination containing ITS sequence (individually and in combination), four species (including *A. crenata*, *A. crispa*, *A. corymbifera* and *A. elegans*) were completely indistinguishable, while the other species were identified to different degrees (**Figures 2, 3**). In NJ tree based on ITS marker, A. quinquegona could not be identified at the species level (**Figure 2A**). In NJ trees based on ITS + psbA-tmH, ITS + psbA-tmH + rbcL and ITS + matK + psbA-tmH + rbcL, A. corymbifera *var.* tuberifera could not be identified (**Figures 2B, 3A,D**). In the NJ tree based on ITS + matK, *A. omissa* could not be identified (**Figure 2C**). In the NJ tree based on ITS + rbcL, A. corymbifera var. Tuberifera and *A. omissa* could not be identified (**Figure 2D**). In the NJ tree based on ITS + matK + psbA-tmH, A. corymbifera var. tuberifera and *A. mamillata* could not be identified (**Figure 3B**). In the NJ tree based on ITS + matK + rbcL, A. corymbifera var. tuberifera and *A. omissa* could not be identified (**Figure 3C**). These results suggest that the different combinations of ITS and the other three chloroplast markers could only increase the species discrimination ability for a few species. ITS was the most suitable barcode for species identification in Chinese Ardisia, and chloroplast markers were supplementary barcodes. The results of the local BLAST method were consistent with those of the NJ method (**Table 2**).

Taxon DNA Method to Analyze the Identification Result

The Taxon DNA software used two standards, namely "Best match" and "Best close match," to evaluate the species identification rate of four single gene fragments and 11 combinations of multiple gene fragments (**Supplementary Table 3**). The ITS fragment yielded the highest rate of species identification, successfully identified 115 samples at the species level, with a success rate of 95.04%. The species identification rate of the three cpDNA fragments and their combination were low. The success rates of species identification of the ITS fragment and three cpDNA combined fragments could not be significantly improved. The results were consistent with the analyses based on the NJ tree and the local BLAST method. The identification rate of the ITS fragment in Chinese Ardisia species was the highest. Hence, ITS could be used as the standard barcode for the genus Ardisia. Finally, the cpDNA gene fragments were used to compensate when individual species could not be identified by ITS.

DISCUSSION

The Efficiency of PCR and Sequencing

A fundamental function of DNA barcode is to determine appropriate barcode sequences. After comprehensive screenings of gene regions of the plant genome, one nuclear (ITS) gene and three chloroplast genes (*rbc*L, *mat*K, and *trn*H-*psb*A) regions have been considered the core barcode in most plants (China Plant BOL Group, 2011; Hollingsworth et al., 2011). The efficiency of PCR amplification and the success rate of sequencing are important indicators for evaluating DNA barcode. In this study,

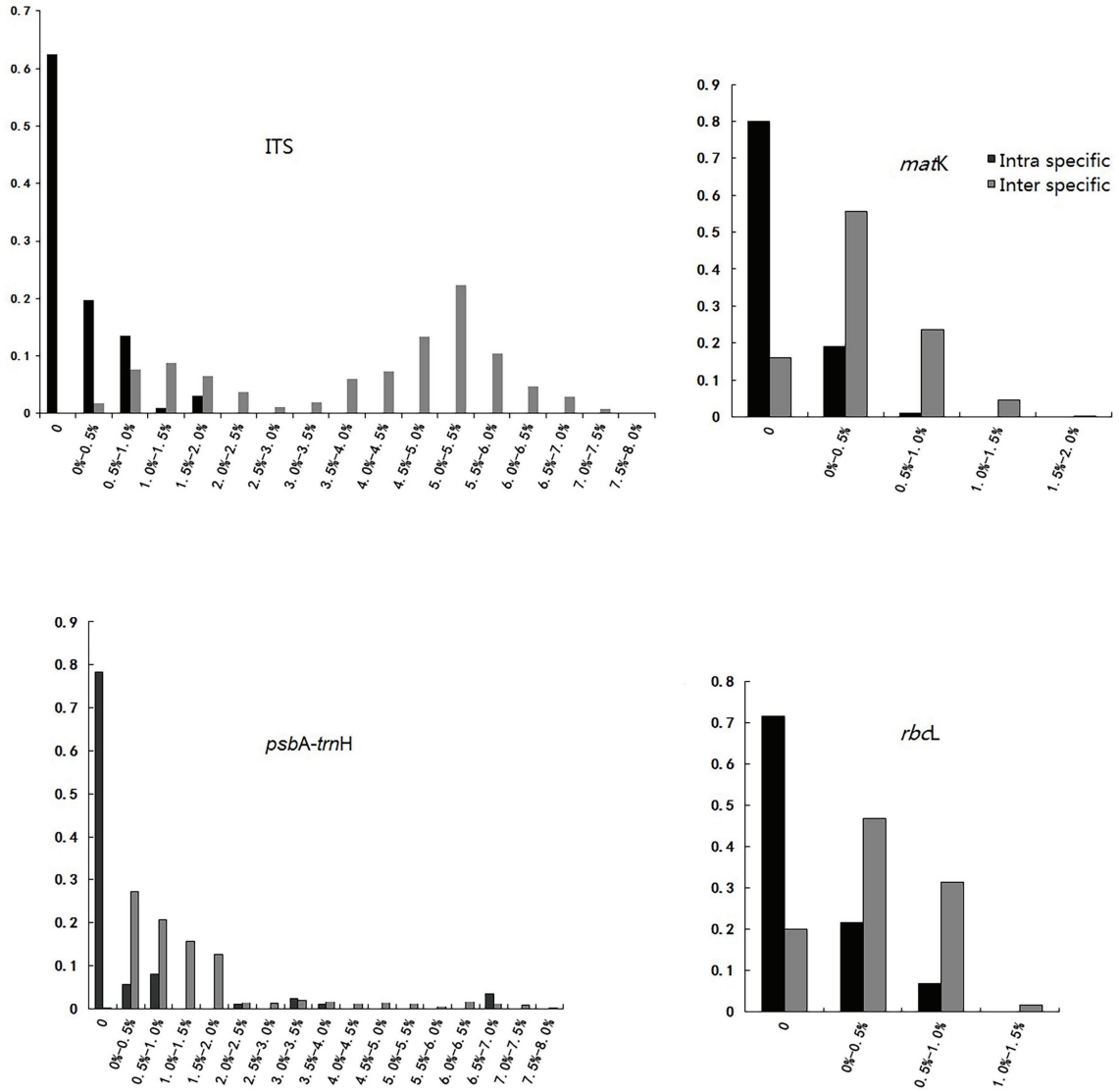

FIGURE 1 | "Barcoding Gap" of each single candidate barcode.

the efficiency of PCR amplification, from high to low, was 100.0% (ITS), 98.35% (*mat*K), 94.21% (*rbc*L), and 94.21% (*psb*A-*trn*H). Three cpDNA fragments could not be successfully amplified in *A. arborescens*, *A. omissa*, and *A. obtusa* samples. The result might be caused by poor DNA template quality or any nucleotide variation in the primer binding region of the species.

Species Identification Efficiency Using Three Chloroplast Regions

Plants of *Ardisia* were used as medicinal plants in China for a long time, but their similar morphological characteristics made them very difficult to differentiate. Liu et al. (2013) analyzed four markers (*psb*A-*trn*H, ITS2, *rbc*L, and matK) in 54 samples representing 24 species of the genus *Ardisia* to choose suitable DNA markers for authenticating and differentiating *Ardisia* at species level. But the sample numbers were limited to research, and most species had no repeated sampling. A small sample size of species may lead to underestimating intraspecific variation or overestimating interspecific differences without analysis of sister groups, which resulted in high validity and accuracy of DNA barcode identification in DNA barcode analysis results (Zhang et al., 2012; Yan et al., 2015). In this study, to identify suitable DNA markers, the sample size was increased to 121 samples representing 33 *Ardisia* species from China, with most species having two or three replicate samples.

The *rbc*L fragment is easily amplified and sequenced, and was chosen to identify flowering plants by Kress et al. (2005). The amplification rate of the *rbc*L primer for Chinese *Ardisia* species was almost 94% in this study, but the species identification rates in the three analysis methods were less than 20%. This result was consistent with the previous results according to which

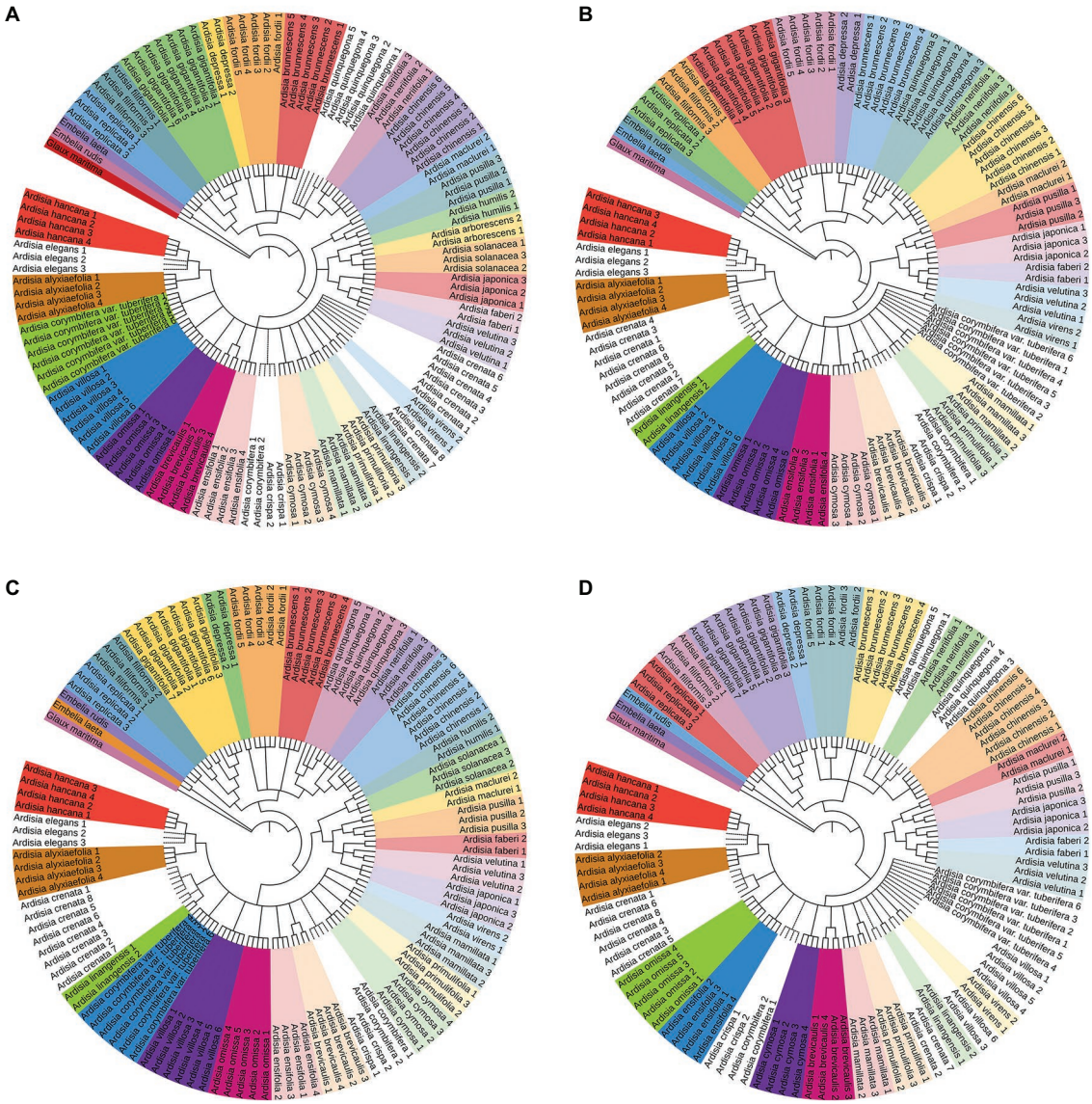

FIGURE 2 | The NJ tree based on four sequences. Successfully identified species are with bootstrap values above 60%. The dotted line indicates unsuccessful identified species. **(A)** The NJ tree based on ITS sequences, **(B)** the NJ tree based on ITS + *psb*A-*trn*H sequences, **(C)** the NJ tree based on ITS + *mat*K sequences, **(D)** the NJ tree based on ITS + *rbc*L sequences.

*rbc*L showed the lowest discrimination rate in seed plants (Lahaye et al., 2008; Ning et al., 2008). The *rbc*L sequences were not suitable as the DNA barcode sequences in Chinese *Ardisia*.

Non-coding regions have their advantages in the study of barcodes. For example, *psb*A-*trn*H is one of the fastest evolving segments in chloroplast fragments and is convenient for primer design (Hollingsworth et al., 2011). Although the length variation of *psb*A-*trn*H sequences was large, the species identification rate of this fragment in the NJ tree and local Blast method was less than 30%. Thus, *psb*A-*trn*H sequences were not suitable as the DNA barcode sequences in Chinese *Ardisia*.

Previous studies suggested that the species discrimination rate would drop significantly when *mat*K sequences were used to distinguish between closely related species or an extensively sampled genus (Ren et al., 2010; Yan et al., 2011). When 54 samples of 24 species from the genus *Ardisia* were analyzed, the results indicated that the *mat*K region was a promising DNA barcode, with the highest species identification efficiency at 91.7% by the nearest distance method and 98.1% using the basic local alignment search tool method. When 121 samples of 33 Chinese *Ardisia* species were used in our study, the amplification and sequencing success rate of *mat*K sequences for Chinese *Ardisia* species was 98.35%. Still, species identification rates obtained through the three analysis methods were less than 50%. However, the performance of *mat*K was overestimated when insufficient samples were used. The discrimination ability of *mat*K for Chinese

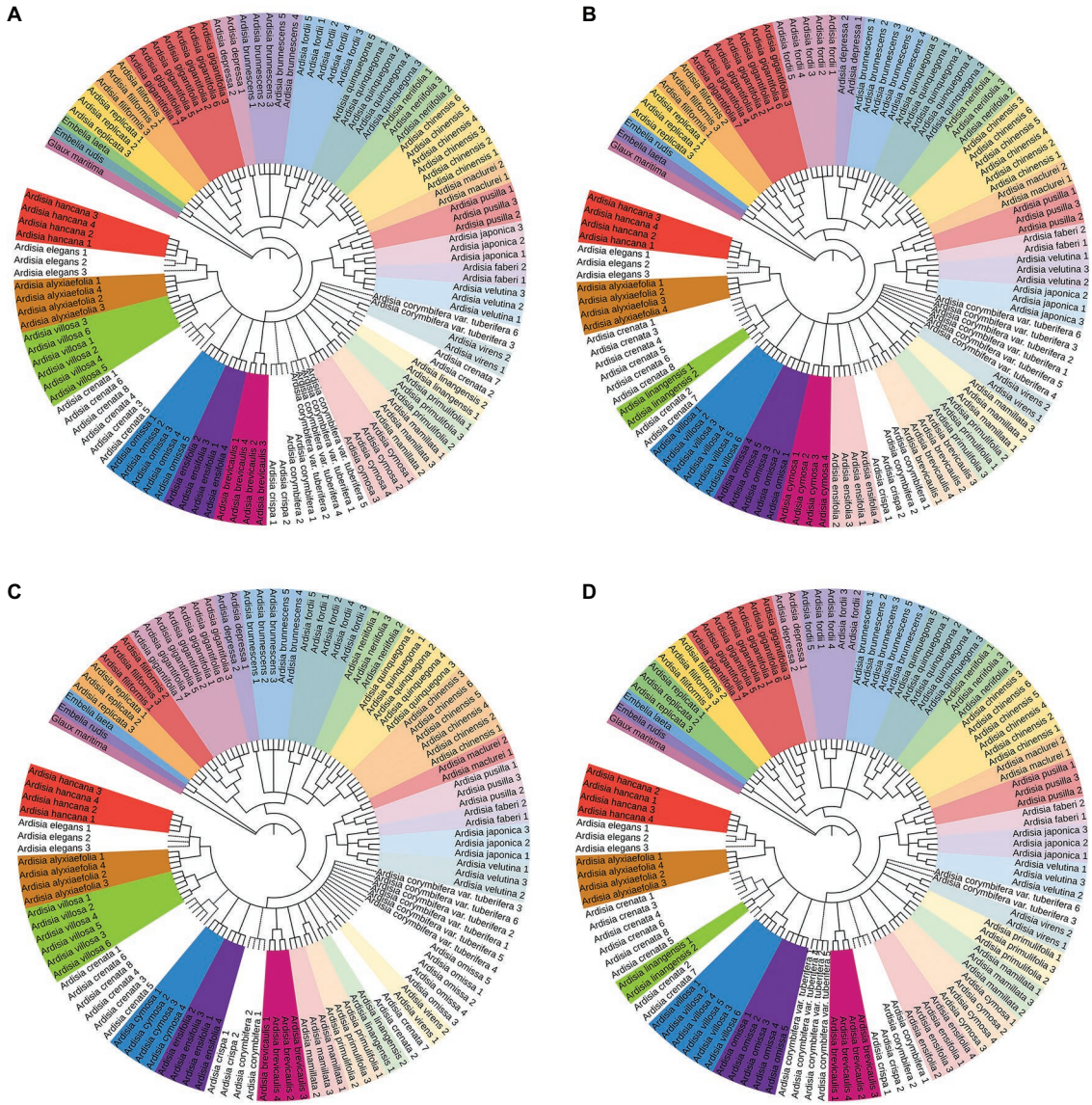

FIGURE 3 | The NJ tree based on other four sequences. Successfully identified species are with bootstrap values above 60%. The dotted line indicates unsuccessful identified species. **(A)** The NJ tree based on ITS + *psb*A-*trn*H + *rbc*L sequences, **(B)** the NJ tree based on ITS + *mat*K + *psb*A-*trn*H sequences, **(C)** the NJ tree based on ITS + *mat*K + *rbc*L sequences, **(D)** the NJ tree based on ITS + *psb*A-*trn*H + *mat*K + *rbc*L sequences.

Ardisia was low, and it was also not suitable as a barcode for the identification of Chinese *Ardisia* species.

When the identification rate of a single fragment failed to satisfy the requisites of their purpose, DNA barcode combinations could improve the species identification efficiency (CBOL Plant Working Group, 2009; Li et al., 2015; Yan et al., 2015). In this study, the effects of using a combination of chloroplast fragments gave limited improvement, consistent with the results in Lysimachia (Myrsinaceae) and Codonopsis (Campanulaceae) (Zhang et al., 2012; Wang et al., 2017). The species identification rates using the three cpDNA fragment combinations were less than 50%. Therefore, they are not suitable DNA barcodes for Chinese *Ardisia* species identification.

Species Identification Efficiency of ITS

Internal transcribed spacer exhibited high species resolution and was proposed as the core barcode for seed plants (Li et al., 2015). While the cpDNA sequences in 5–10% of angiosperms were not easily amplified, the ITS sequences could be amplified more easily. This study obtained the same result that the amplification success rate of ITS was high, reaching 100%. In flowering plants, ITS sequences perform efficiently as molecular markers, with 3–4 times higher variation sites than cpDNA fragments (Zhang et al., 2012). Our results were consistent with those of previous studies. Compared to *mat*K, *rbc*L, and *psb*A-*trn*H the number of variation sites of ITS was almost four, nine, and four times higher, respectively.

TABLE 2 | Identification success rates obtained using NJ tree and local blast analysis methods for each single candidate barcodes and combinations of them.

Single candidate barcodes and combinations of them	NJ tree	Similarity-based
	Method*	Method (BLAST)
ITS	84.85% (28/33)	84.85% (28/33)
*mat*K	31.25% (10/32)	31.25% (10/32)
*psb*A-*tm*H	33.33% (10/30)	33.33% (10/30)
*rbc*L	16.67% (5/30)	16.67% (5/30)
ITS + *mat*K	84.38% (27/32)	84.38% (27/32)
ITS + *rbc*L	80.00% (24/30)	80.00% (24/30)
ITS + *psb*A-*tm*H	83.33% (25/30)	83.33% (25/30)
*mat*K + *psb*A-*tm*H	46.67% (14/30)	46.67% (14/30)
*mat*K + *rbc*L	33.33% (10/30)	33.33% (10/30)
*psb*A-*tm*H + *rbc*L	33.33% (10/30)	33.33% (10/30)
ITS + *mat*K + *psb*A-*tm*H	80.00% (24/30)	80.00% (24/30)
ITS + *mat*K + *rbc*L	80.00% (24/30)	80.00% (24/30)
ITS + *psb*A-*tm*H + *rbc*L	83.33% (25/30)	83.33% (25/30)
*mat*K + *psb*A-*tm*H + *rbc*L	43.33% (13/30)	43.33% (13/30)
ITS + *mat*K + *psb*A-*tm*H + *rbc*L	83.33% (25/30)	83.33% (25/30)

**Based on the proportion of monophyletic species with > 60% bootstrapping.*

Fu et al. (2011) demonstrated that the identification rate of Vitaceae species was 93.8%. Wilcoxon signed-rank tests reveal that ITS was stronger than other core barcodes regardless of interspecific or intraspecific variations. Our results also suggested that the ITS fragment's identification rate using three analysis methods in Chinese *Ardisia* species was higher than 80%, but closer to 100%. ITS showed the highest number of variation sites and the most efficient amplification, sequencing, and identification among the core barcodes. After combining three cpDNA fragments with ITS, the corresponding rate of species identification did not increase significantly using three analysis methods. Therefore, considering cost and experimental effectiveness, we recommend only ITS to identify Chinese *Ardisia* plants.

Classification of Chinese *Ardisia* Species

Except for *A. crispa* and *A. corymbifera*, the other species were completely resolved by single or combination markers. Because the sequences of the two species are identical, samples of *A. crispa* and *A. corymbifera* formed a monophyletic clade in the NJ trees (**Figures 2, 3**), and these two species were not resolved. Other DNA barcoding loci should further analyze the unresolved species.

Samples of *A. crenata* and *A. linangensis*, formed a monophyletic clade in NJ trees based on ITS + *matK*, ITS + *psb*A-*tm*H, ITS + *mat*K + *psb*A-*tm*H and ITS + *mat*K + *psb*A-*tm*H + *rbc*L combination sequences. Two *A. linangensis* samples always formed a monophyletic clade in the NJ tree analysis. The results were consistent with the findings of Wang and Xia (2012, 2013). Therefore, we also did not support the idea that *A. crenata* var. *bicolor* should be merged with A. crenata to form the *A. crenata* complex. However, the samples were limited, and more samples should be collected and analyzed to determine the *A. crenata* complex.

CONCLUSION

Ardisia plants have been mostly used for medicinal purposes in China for a long time, with good effects on rheumatism, phthisis and various kinds of inflammation. Due to the high degree of similarity in morphology and the phenomenon of homonymy, the traditional identification methods can not accurately identify the species. For the sake of drug safety, the species identification ability of four DNA barcodes (*psb*A-*trn*H, ITS, *rbc*L, *mat*K) in *Ardisia* plants in China was evaluated in this study. The results showed that ITS showed the highest number of variation sites and the most efficient amplification, sequencing and identification. Using three cpDNA fragments alone, combined with each other or combined with ITS fragments, the efficiency of species identification did not significantly increase. Considering the cost and experimental effect, we recommend ITS as the core barcode for plant identification of *Ardisia* in China.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

ZL and BZ supervised the whole project. CX and WS performed the major research and wrote the manuscript in equal contribution. LW provided the technical support and language editing support. RX, YZ, and WZ provided their professional expertise. HJ and Panjwani provided data analysis. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by National Key R&D Program of China from the Ministry of Science and Technology of China (No. 2021YFE0100900), the National Science Foundation of China (No. 81903758), and the Fundamental Research Funds for the Central public welfare research institutes (ZZ13-YQ-106).

ACKNOWLEDGMENTS

We thank the Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences (ICMM), Guilin Botany, Guangxi Institute Botany, Chinese Academy of sciences. We thank Shizhong Mao for specimen identification, and thank Jingjian Li for providing samples.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.860778/full#supplementary-material

REFERENCES

Amritha, N., Bhooma, V., and Parani, M. (2020). Authentication of the market samples of Ashwagandha by DNA barcoding reveals that powders are significantly more adulterated than roots. *J. Ethnopharmacol.* 256:112725. doi: 10.1016/j.jep.2020.112725

CBOL Plant Working Group (2009). A DNA barcode for land plants. *Proc. Natl. Acad. Sci. U. S. A.* 106, 12794–12797. doi: 10.1073/pnas.0905845106

Chen, S. L. (2015). *Standard DNA Barcodes of Chinese Materia Medica in Chinese Pharmacopoeia*. Beijing: Medical Science and Technology Press.

Chen, S. L., Pang, X. H., Song, J. Y., Shi, L. C., Yao, H., Han, J. P., et al. (2014). A renaissance in herbal medicine identification: from morphology to DNA. *Biotechnol. Adv.* 32, 1237–1244. doi: 10.1016/j.biotechadv.2014.07.004

Chen, J., and Pipoly, J. J. (1996). "Myrsinaceae," in *Flora of China*. eds. Z. Wu and P. H. Raven (Beijing: Science Press), 1–38.

Chen, S. L., Yao, H., Han, J. P., Liu, C., Song, J. Y., Shi, L. C., et al. (2010). Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PLoS One* 5:e8613. doi: 10.1371/journal.pone.0008613

China Plant BOL Group, Li, D. Z., Gao, L. M., Li, H. T., Wang, H., Ge, X. J., et al. (2011). Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proc. Natl. Acad. Sci. U. S. A.* 108, 19641–19646. doi: 10.1073/pnas.1104551108

Cui, X., Li, W., Wei, J., Qi, Y., and Zheng, X. (2020). Assessing the identity of commercial herbs from a Cambodian market using DNA barcoding. *Front. Pharmacol.* 11:244. doi: 10.3389/fphar.2020.00244

Dai, W., Dong, P., Tian, S., and Mei, Q. (2018). A pharmacognostical study on *Ardisia gigantifolia* and its adulterants. *Med. Plant.* 9, 39–44. doi: 10.19600/j.cnki.issn2152-3924.2018.04.011

Fazekas, A. J., Burgess, K. S., Kesanakurti, P. R., Graham, S. W., Newmaster, S. G., Husband, B. C., et al. (2008). Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. *PLoS One* 3:e2802. doi: 10.1371/journal.pone.0002802

François, C., Hul, S., Deharo, E., and Bourdy, G. (2016). Natural remedies used by Bunong people in Mondulkiri province (Northeast Cambodia) with special reference to the treatment of 11 most common ailments. *J. Ethnopharmacol.* 191, 41–70. doi: 10.1016/j.jep.2016.06.003

Fu, Y. M., Jiang, W. M., and Fu, C. X. (2011). Identification of species within *Tetrastigma* (Miq.) Planch. (Vitaceae) based on DNA barcoding techniques. *J. Syst. Evol.* 49, 237–245. doi: 10.1111/j.1759-6831.2011.00126.x

Hebert, P. D. N., Cywinska, A., Ball, S. L., and de Waard, J. R. (2003a). Biological identification through DNA barcodes. *P. Roy. Soc. B-Biol. Sci.* 270, 313–321. doi: 10.1098/rspb.2002.2218

Hebert, P. D. N., Ratnasingham, S., and de Waard, J. R. (2003b). Barcoding animal life: cytochrome *c* oxidase subunit 1 divergences among closely related species. *P. Roy. Soc. B-Biol. Sci.* 270(Suppl. 1), S96–S99. doi: 10.1098/rsbl.2003.0025

Hollingsworth, P. M., Graham, S. W., and Little, D. P. (2011). Choosing and using a plant DNA barcode. *PLoS One* 6:e19254. doi: 10.1371/journal.pone.0019254

Joaquín-Cruz, E., Dueñas, M., García-Cruz, L., Salinas-Moreno, Y., Santos-Buelga, C., and García-Salinas, C. (2015). Anthocyanin and phenolic characterization, chemical composition and antioxidant activity of chagalapoli (*Ardisia compressa* k.) fruit: a tropical source of natural pigments. *Food Res. Int.* 70, 151–157. doi: 10.1016/j.foodres.2015.01.033

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitution through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16, 111–120. doi: 10.1007/BF01731581

Kobayashi, H., and de Mejía, E. (2005). The genus Ardisia: a novel source of health-promoting compounds and phytopharmaceuticals. *J. Ethnopharmacol.* 96, 347–354. doi: 10.1016/j.jep.2004.09.037

Kress, W. J., Wurdack, K. J., Zimmer, E. A., Weigt, L. A., and Janzen, D. H. (2005). Use of DNA barcodes to identify flowering plants. *Proc. Natl. Acad. Sci. U. S. A.* 102, 8369–8374. doi: 10.1073/pnas.0503123102

Lahaye, R., Van der, B. M., Bogarin, D., Warner, J., Pupulin, F., Gigot, G., et al. (2008). DNA barcoding the floras of biodiversity hotspots. *Proc. Natl. Acad. Sci. U. S. A.* 105, 2923–2928. doi: 10.1073/pnas.0709936105

Larkin, M. A., Blackshields, G., Brown, N., 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

Li, D. Z., Gao, L. M., Li, H. T., Wang, H., Ge, X. J., Liu, J. Q., et al. (2011). Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proc. Natl. Acad. Sci. U. S. A.* 108, 19641–19646. doi: 10.1073/pnas.1104551108

Li, X., Yang, Y., Henry, R. J., Rossetto, M., Wang, Y., and Chen, S. (2015). Plant DNA barcoding: from gene to genome. *Biol. Rev.* 90, 157–166. doi: 10.1111/brv.12104

Liu, Y. M., Wang, K., Liu, Z., Luo, K., Chen, S. L., and Chen, K. L. (2013). Identification of medical plants of 24 *Ardisia* species from China using the *mat* K genetic marker. *Pharmacogn. Mag.* 9, 331–337. doi: 10.4103/0973-1296.117829

Meier, R., Shiyang, K., Vaidya, G., and Ng, P. K. L. (2006). DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Syst. Biol.* 55, 715–728. doi: 10.1080/10635150600969864

Meyer, C. P., and Paulay, G. (2005). DNA barcoding: error rates based on comprehensive sampling. *PLoS Biol.* 3:e422. doi: 10.1371/journal.pbio.0030422

Newell, A. M. B., Yousef, G. G., Lila, M. A., Ramírez-Mares, M. V., and de Mejia, E. G. (2010). Comparative in vitro bioactivities of tea extracts from six species of *Ardisia* and their effect on growth inhibition of hepg2 cells. *J. Ethnopharmacol.* 130, 536–544. doi: 10.1016/j.jep.2010.05.051

Ning, S. P., Yan, H. F., Hao, G., and Ge, X. J. (2008). Current advances of DNA barcoding study in plants. *Biodivers. Sci.* 16, 417–425. doi: 10.3724/SP.J.1003.2008.08215

Oliveira, N. A., Santos, O. Y. I., Lima, A. B., Elisabete, A. D. M. M., and Frota, A. G. (2018). Pharmacological effects of the isomeric mixture of alpha and beta amyrin from protium heptaphyllum: a literature review. *Fund. Clin. Pharmacol.* 33, 4–12. doi: 10.1111/fcp.12402

Poudel, R. C., Li, D. Z., and Forrest, A. (2011). High universality of matK primers for barcoding gymnosperms. *J. Syst. Evol.* 49, 169–175. doi: 10.1111/j.1759-6831.2011.00128.x

Priyanka, M., Amit, K., Akshitha, N., Daya, N. M., Ashutosh, S., Rakesh, T., et al. (2015). DNA barcoding: an efficient tool to overcome authentication challenges in the herbal market. *Plant Biotechnol. J.* 14, 8–21. doi: 10.1111/pbi.12419

Ren, B. Q., Xiang, X. G., and Chen, Z. D. (2010). Species identification of *Alnus* (Betulaceae) using nrDNA and cpDNA genetic markers. *Mol. Ecol. Resour.* 10, 594–605. doi: 10.1111/j.1755-0998.2009.02815.x

Sanjeev, S., Murthy, M. K., Devi, M. S., Khushboo, M., Renthlei, Z., Ibrahim, K. S., et al. (2019). Isolation, characterization, and therapeutic activity of bergenin from marlberry (*Ardisia colorata* Roxb.) leaf on diabetic testicular complications in Wistar albino rats. *Environ. Sci. Pollut. R.* 26, 7082–7101. doi: 10.1007/s11356-019-04139-9

Santhosh Kumar, J. U., Krishna, V., Seethapathy, G. S., Senthilkumar, U., Ragupathy, S., Ganeshaia, K. N., et al. (2015). DNA barcoding to assess species adulteration in raw drug trade of "bala" (genus: sida l.) herbal products in South India. *Biochem. Syst. Ecol.* 61, 501–509. doi: 10.1016/j.bse.2015.07.024

Skjua, B., Mr, C., Gss, D., Vk, A., Rus, B., and Gr, E. (2020). DNA barcoding of momordica species and assessment of adulteration in momordica herbal products, an anti-diabetic drug. *Plant Gene* 22:100227. doi: 10.1016/j.plgene.2020.100227

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* 28, 2731–2739. doi: 10.1093/molbev/msr121

Tao, T. (2010). Standalone BLAST setup for windows PC. Available at: http://www.ncbi.nlm.nih.gov/books/NBK52637/ (Accessed August 31, 2020).

Voravuthikunchai, S., Lortheeranuwat, A., Jeeju, W., Sririrak, T., Phongpaichit, S., and Supawita, T. (2004). Effective medicinal plants against enterohaemorrhagic escherichia coli o157:h7. *J. Ethnopharmacol.* 94, 49–54. doi: 10.1016/j.jep.2004.03.036

Wang, D. Y., Wang, Q., Wang, Y. L., Xiang, X. G., Huang, L. Q., and Jin, X. H. (2017). Evaluation of DNA barcodes in *Codonopsis* (Campanulaceae) and in some large angiosperm plant genera. *PLoS One* 12:e0170286. doi: 10.1371/journal.pone.0170286

Wang, J., and Xia, N. H. (2012). *Ardisia crenata Complex (Primulaceae) Studies Using Morphological and Molecular Data*. ed. J. K. Mworia (Botany, London: InTech), 163–172.

Wang, J., and Xia, N. H. (2013). Quantitative analysis of morphological characters of *Ardisia crenata* complex (Primulaceae). *J. Trop. Subtrop. Bot.* 6, 543–548.

Xin, X., Yu, D., Zhu, L., Gu, Z. X., Yuan, L., and Huang, S. (2015). Qualitative and quantitative method for compound Aidicha tablets. *Cent. South Pharma.* 13, 410–413.

Yan, H. F., Hao, G., Hu, C. M., and Ge, X. J. (2011). DNA barcoding in closely related species: a case study of *Primula* L. sect. *Proliferae* Pax (Primulaceae) in China. *J. Syst. Evol.* 49, 225–236. doi: 10.1111/j.1759-6831.2011.00115.x

Yan, L. J., Liu, J., Möller, M., Zhang, L., Zhang, X. M., Li, D. Z., et al. (2015). DNA barcoding of *rhododendron* (Ericaceae), the largest chinese plant genus in biodiversity hotspots of the himalaya–hengduan mountains. *Mol. Ecol. Resour.* 15, 932–944. doi: 10.1111/1755-0998.12353

Yang, C. Q., Lv, Q., and Zhang, A. B. (2020). Sixteen years of DNA barcoding in China: what has been done? What can be done? *Front. Ecol. Evol.* 8:57. doi: 10.3389/fevo.2020.00057

Yao, H., Song, J. Y., Liu, C., Luo, K., Han, J. P., Li, Y., et al. (2010). Use of ITS2 region as the universal DNA barcode for plants and animals. *PLoS One* 5:e13102. doi: 10.1371/journal.pone.0013102

Zhang, D., and Jiang, B. (2020). Species identification in complex groups of medicinal plants based on DNA barcoding: a case study on *Astragalus* spp. (Fabaceae) from Southwest China. *Conserv. Genet. Resour.* 12, 469–478. doi: 10.1007/s12686-019-01119-6

Zhang, C. Y., Wang, F. Y., Yan, H. F., Hao, G., Hu, C. M., and Ge, X. J. (2012). Testing DNA barcoding in closely related groups of *Lysimachia* L. (Myrsinaceae). *Mol. Ecol. Resour.* 12, 98–108. doi: 10.1111/j.1755-0998.2011.03076.x

Zhao, C. Q., Zhou, Y., Ping, J., and Xu, L. M. (2014). Traditional Chinese medicine for treatment of liver diseases: progress, challenges and opportunities. *J. Integr. Med.* 12, 401–408. doi: 10.1016/S2095-4964(14)60039-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.

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

Copyright © 2022 Xiong, Sun, Wu, Xu, Zhang, Zhu, J., Panjwani, Liu and Zhao. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 15 April 2022
doi: 10.3389/fpls.2022.873788

Nucleotide Evolution, Domestication Selection, and Genetic Relationships of Chloroplast Genomes in the Economically Important Crop Genus *Gossypium*

Tong Zhou[1†], Ning Wang[1†], Yuan Wang[1], Xian-Liang Zhang[2], Bao-Guo Li[1], Wei Li[2], Jun-Ji Su[3], Cai-Xiang Wang[3], Ai Zhang[3], Xiong-Feng Ma[2] and Zhong-Hu Li[1*]*

[1] Shaanxi Key Laboratory for Animal Conservation, Key Laboratory of Resource Biology and Biotechnology in Western China (Ministry of Education), College of Life Sciences, Northwest University, Xi'an, China, [2] State Key Laboratory of Cotton Biology, Institute of Cotton Research, Chinese Academy of Agricultural Sciences, Anyang, China, [3] Gansu Provincial Key Laboratory of Aridland Crop Science, College of Life Science and Technology, Gansu Agricultural University, Lanzhou, China

Gossypium hirsutum (upland cotton) is one of the most economically important crops worldwide, which has experienced the long terms of evolution and domestication process from wild species to cultivated accessions. However, nucleotide evolution, domestication selection, and the genetic relationship of cotton species remain largely to be studied. In this study, we used chloroplast genome sequences to determine the evolutionary rate, domestication selection, and genetic relationships of 72 cotton genotypes (36 cultivated cotton accessions, seven semi-wild races of *G. hirsutum*, and 29 wild species). Evolutionary analysis showed that the cultivated tetraploid cotton genotypes clustered into a single clade, which also formed a larger lineage with the semi-wild races. Substitution rate analysis demonstrated that the rates of nucleotide substitution and indel variation were higher for the wild species than the semi-wild and cultivated tetraploid lineages. Selection pressure analysis showed that the wild species might have experienced greater selection pressure, whereas the cultivated cotton genotypes underwent artificial and domestication selection. Population clustering analysis indicated that the cultivated cotton accessions and semi-wild races have existed the obviously genetic differentiation. The nucleotide diversity was higher in the semi-wild races compared with the cultivated genotypes. In addition, genetic introgression and gene flow occurred between the cultivated tetraploid cotton and semi-wild genotypes, but mainly *via* historical rather than contemporary gene flow. These results provide novel molecular mechanisms insights into the evolution and domestication of economically important crop cotton species.

Keywords: cotton, domestication selection, gene flow, genetic relationship, nucleotide evolution

OPEN ACCESS

Edited by:
Lin-Feng Li,
Fudan University, China

Reviewed by:
Jie Qiu,
Shanghai Normal University, China
Nian Wang,
Huazhong Agricultural University, China
Xiongming Du,
State Key Laboratory of Cotton Biology, Cotton Institute of the Chinese Academy of Agricultural Sciences, China

***Correspondence:**
Zhong-Hu Li
lizhonghu@nwu.edu.cn
Xiong-Feng Ma
maxf_caas@163.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: *11 February 2022*
Accepted: *24 March 2022*
Published: *15 April 2022*

Citation:
Zhou T, Wang N, Wang Y, Zhang X-L, Li B-G, Li W, Su J-J, Wang C-X, Zhang A, Ma X-F and Li Z-H (2022) Nucleotide Evolution, Domestication Selection, and Genetic Relationships of Chloroplast Genomes in the Economically Important Crop Genus Gossypium. Front. Plant Sci. 13:873788. doi: 10.3389/fpls.2022.873788

INTRODUCTION

Since Darwin's time, biologists have recognized that investigating the human domestication of wild plants can help to improve our understanding of the evolutionary process (Yoo et al., 2014). Generally, domesticated forms of cultivated species differ from their wild counterparts in numerous traits (Hu et al., 2013; Mabry et al., 2021). Insights into the evolution of chloroplast genome's domestication and selection are made possible by comparative studies of wild and domesticated representatives of individual cultivated species. In the previous study, scholars used chloroplast genome data to analyze the genetic variation and evolution of olive. As a control, the cultivated species were employed to analyze genome variation and genetic association among olive chloroplasts (Niu et al., 2020). Meanwhile, some other study have also examined the evolutionary mechanism of the chloroplast genome of cultivated *Camellia sinensis* and its relatives (Li et al., 2021). In recent studies, comparisons of wild and domesticated plants have provided important insights into the developmental mechanisms that underlie traits affected strongly due to targeted selection by humans (Yoo et al., 2014). In general, domesticated plants are characterized by reduced genetic variation and relaxed selection pressure compared with their wild counterparts. Several studies also found high levels of continuous gene flow from wild to cultivated genotypes (Price, 2002; Burger et al., 2008; Gross and Olsen, 2010; Ma et al., 2019). Thus, the domestication process may provide a basis for studying the overall evolutionary relationships associated with wild crop transformation and identifying the genes under selection (Gepts, 2004; Burger et al., 2008).

Cotton (*Gossypium*) is one of the most important crops worldwide (Wendel, 1989; Ruan, 2003) and a major source of natural fiber for the textile industry. Allopolyploid cotton originated in the New World and diverged into at least six species throughout the tropical and subtropical Americas: *G. hirsutum* (AD_1), *G. barbadense* (AD_2), *G. tomentosum* Nuttalex Seemann (AD_3), *G. mustelinum* Miersex Watt (AD_4), *G. darwinii* Watt (AD_5), and *G. ekmanianum* (AD_6) (Wendel and Cronn, 2003; Wendel and Grover, 2015). The diploid species comprise eight monophyletic genome groups: A, B, C, D, E, F, G, and K (Wendel and Cronn, 2003; Grover et al., 2007; Wendel et al., 2010). These groups can be separated into three main lineages in three continental regions: 13 D-genome species from the American continents, 15 species from the Asian and African continents (A-, B-, E-, and F-genomes), and 18 species (C-, G- and K-genomes) from Australia (Wendel and Cronn, 2003). Hence, cotton species provide a fascinating model system for studying evolution, domestication selection, genetic introgression, and gene flow among different continents (Fryxell, 1969, 1978; Wendel, 1989; Wendel and Grover, 2015; Chen et al., 2016, 2017a,b). Four species in the genus *Gossypium* are cultivated for the production of spinnable fiber, i.e., two allotetraploid species comprising *G. hirsutum* L. and *G. barbadense* L. (2n = 4x = 52), and two diploid species comprising *Gossypium herbaceum* L. (A_1) and *Gossypium arboretum* L. (A_2) (2n = 2x = 26) (Wendel and Cronn, 2003; Wendel and Grover, 2015). Allopolyploid cottons were considered to be about 1.5 million years old and were domesticated by humans 4,000 to 5,000 years ago (Wendel, 1989; Wang et al., 2017), which were originally domesticated from tree cotton in the Mesoamerican and Caribbean regions, and then further domesticated and improved in the southern United States (Fang et al., 2017). And two diploid cotton species, *G. arboretum* and *G. herbaceum*, have been cultivated for several millennia (Simon et al., 2016), which were initially domesticated on Madagascar or in the Indus Valley (Mohenjo Daro), and was subsequently dispersed to Africa and other areas of Asia (Wendel and Grover, 2015; Du et al., 2018; Huang et al., 2020). Due to the high-yield characteristics of allopolyploid cottons, the American upland cottons have been introduced and replaced by two diploid cotton species (*G. arboretum* and *G. herbaceum*) (Fang et al., 2017; Du et al., 2018). Up to now, the Upland cotton (*G. hirsutum*) accounts for more than 95% of the worldwide production of cotton (Yoo et al., 2014; Fang et al., 2017; Ma et al., 2019; Wang et al., 2019; Zhang et al., 2020).

Following human-mediated selection and agronomic improvement, the ability of cotton species to adapt to various environments was enhanced and the production of fiber from cotton improved significantly (Ma et al., 2019). The domestication process also resulted in other morphological changes in other crops such as sorghum, rice and soybean (Ma et al., 2019), including early flowering, larger and/or more fruits, annualized habit, plant height reduction, and loss of seed dormancy (Yoo et al., 2014). When plants undergo artificial domestication, the relaxation of certain features is inevitable (Price, 2002), that is, when plants undergo relatively large changes, such as from the transition from nature to domestication, certain characteristics important for survival in nature lose much of their adaptive significance under artificial directional selection. Hence, one would expect natural selection for such characteristics to lose its intensity (Coss, 1999; Price, 2002). Many studies have shown that the genetic diversity of upland cotton varieties is low, mainly due to several bottlenecks in the domestication process (Brubaker and Wendel, 1994; May et al., 1995; Iqbal et al., 2001; Wendel and Cronn, 2003). In addition, previous studies based on whole-genome resequencing of upland cotton have indicated that the genomic diversity of upland cotton decreased under the stress of artificial selection (Fang et al., 2017; Ma et al., 2019). Thus, in the current era of genomic big data, high-throughput "omics" sequencing techniques allow detailed analyses of the genetic changes associated with artificial domestication, as well as providing new, accurate, and targeted genome-based crop breeding strategies (Wang et al., 2017; Li et al., 2020; Yang et al., 2020). For example, in maize and rice, the use of high-quality backbone parents can obtain notable improvements in breeding efficiency (Ma et al., 2019). The whole-genome sequences of allotetraploid cotton and its ancestors have been completed, and the high-quality allotetraploid upland cotton genome is an effective tool for systematically exploring the genomic mysteries of polyploidy (Li et al., 2014, 2015; Zhang et al., 2015). Compared with whole-genome sequencing, the chloroplast genome is single-copy, maternally inherited, and there is no chain exchange or free combination phenomenon. It has a relatively independent evolutionary route. In addition,

the highly conserved characteristics of the chloroplast genome make them useful for the rapid analysis of species evolution (Jansen et al., 2007; Parks et al., 2009; Wang et al., 2013; Chen et al., 2014). However, the whole-genome resequencing (WGR) is parental inheritance, and there may be genetic recombination (Gover et al., 2020; Wang et al., 2020).

In the current study, to better understand the evolution, domestication selection, and genetic relationships of cotton, we analyzed the chloroplast genomic variation in 72 cotton genotypes comprising *G. hirsutum* and its 29 cultivated upland cotton accessions, *G. barbadense* and its three cultivated accessions (*Gossypium barbadense* cultivar zhonghai 7, *Gossypium barbadense* cultivar Kaiyuan, and *Gossypium barbadense* cultivar yuanmou), *G. africanum*, *G. arboretum*, seven semi-wild races of *G. hirsutum*, and 29 wild cotton species. We also estimated molecular dating, genetic introgression, nucleotide substitutions, and indel variation.

MATERIALS AND METHODS

DNA Extraction and Plant Materials

The fresh leaves of seven semi-wild races of upland cotton, i.e., punctatum, latifolium, richmondi, morrilli, marie-galante, palmeri, and yucatanense, were collected from the National Wild Cotton Nursery in Sanya, China. In addition, 29 cultivated upland cotton accessions were also obtained from different ecological geographic regions, with three accessions from the United States, eight from the Yellow River region, 12 from the Yangtze River area, four from northwest China, and two from north China (**Table 1**). Leaf tissues were dried with silica gel and genomic DNA was extracted using the modified CTAB method (Doyle and Doyle, 1987). Approximately 5 μg of purified DNA was used to construct paired-end libraries with an insert size of 350 bp and sequencing was performed with the Illumina HiSeq 2500 platform by Novogene (Beijing, China). Additionally, we have also downloaded the 36 chloroplast genomes of cotton species from NCBI (National Center for Biotechnology Information) for further combination analysis.

Chloroplast Genome Assembly, and Annotation

The raw sequencing reads obtained by the company (Novogene, Beijing, China) were filtered through the "AmbiguityFiltering.pl" script in the NGSQCToolkit software (Patel and Jain, 2012), and removed the fragments with fuzzy bases greater than 2% and those with bases less than 50 bp. The clean reads were assembled by the MIRA 4.0.2 program (Chevreux et al., 2004) where the complete chloroplast genome of *G. hirsutum* (AD_1) (NC_007944) was used as the reference sequence in this process. In order to further assemble the whole chloroplast genomes, some ambiguous regions were extended using the MITObim v1.7 program with a baiting and iteration method (Hahn et al., 2013). The contigs obtained were used to generate consensus sequences with Geneious v8.0.2 (Kearse et al., 2012). The chloroplast genomes were then annotated using the Dual Organellar Genome Annotator (DOGMA, Wyman et al., 2004) program and manual corrections were made for some specific genes. All tRNA genes were further confirmed using the online tool tRNAscan-SE (Schattner et al., 2005). All of the newly generated genome sequences were submitted to GenBank (accession numbers MK792837–MK792871 and MG800784).

Genetic Clustering Analysis

To evaluate the genetic relationships among cotton genotypes, molecular phylogenetic analysis was conducted using 72 complete chloroplast genome sequences (**Table 1**) and two outgroups comprising *Bombax ceiba* (NC_037494) and *Theobroma cacao* (NC_014676). First, all of the sequences were aligned using the MAFFT program (Katoh and Standley, 2013) and the best-fit model was then selected with Modeltest v3.7 (Posada and Crandall, 1998) based on Akaike's information criterion. Finally, a maximum likelihood tree was constructed using RAxML v7.2.8 (Stamatakis, 2006) where the best model was GTR + G based on 1000 bootstrap replicate tests.

Estimation of Divergence Times

Previously estimated dates of speciation events (fossil records) were used to calibrate the phylogenetic tree (Pfeil and Crisp, 2008). In BEAST v1.8.0 (Drummond et al., 2012), we used the Yule process speciation prior and the uncorrelated lognormal model of rate change with a relaxed clock to estimate the divergence times among cotton lineages. The divergence time was calculated based on 74 chloroplast protein-coding sequences shared by the cotton genotypes, and we used three fossil records: AD_1 (*G. hirsutum*) and A_2 (*G. arboreum*) diverged 1–2 Mya (Wendel, 1989), A_2 (*G. arboreum*) and D_5 (*G. raimondii*) diverged ∼ 5–10 Mya (Senchina et al., 2003), and *Theobroma-Gossypium* diverged 60 Mya (Carvalho et al., 2011). A normal prior probability distribution was used to account for the uncertainty of prior knowledge. The analyses were run for 50,000,000 generations and the parameters were sampled every 5,000 generations. Tracer v 1.6 (Drummond et al., 2012) was used to determine the effective sample size (>200) and the first 20% of the samples were discarded as burn-in. Tree Annotator v.1.8.0 (Drummond et al., 2012) was used to summarize the set of post-burn-in trees and their parameters were used to produce a maximum clade credibility chronogram, which illustrated the mean divergence time estimates in the 95% highest posterior density (HPD) intervals. Finally, FigTree V1.3.1 (Drummond et al., 2012) was used to visualize the molecular dating estimates.

Analysis of Nucleotide Substitutions

Transitions/transversions explain the substitution rates of nucleotides, so we determined the transition/transversion rates using single nucleotide polymorphism (SNP) loci in protein-coding sequences in the cotton chloroplast genome. These analyses were conducted based on two genetic groups obtained from the phylogenetic analyses. One group contained the diploid cotton species (including *G. africanum* and *G. arboretum*) and the other group comprised tetraploid semi-wild races and cultivated upland cotton genotypes (excluding *G. barbadense*). MEGA files generated from SNP data were analyzed with MEGA7 software (Kumar et al., 2016) to obtain the transition/transversion rate.

TABLE 1 | List of taxa sampled in this study and species accession numbers (GenBank).

Number	Species	Accession number	Source	Logogram
1	*Gossypium punctatum*	MK792868	Sanya, China	JBM
2	*Gossypium richmondii*	MK792869	Sanya, China	Iqmd
3	*Gossypium morrilli*	MK792866	Sanya, China	MLE
4	*Gossypium marie-galante*	MK792865	Sanya, China	MLJ
5	*Gossypium palmerii*	MK792867	Sanya, China	PME
6	*Gossypium yucatanense*	MK792870	Sanya, China	YKT1
7	*Gossypium hirsutum* cultivar 06G415	MK792871	Yellow river	S32
8	*Gossypium hirsutum* cultivar antongSP21	MK792837	United States	S24
9	*Gossypium hirsutum* cultivar chuanmian45	MK792838	Yangtze river	S47
10	*Gossypium hirsutum* cultivar CJL-233	MK792839	Yangtze river	S252
11	*Gossypium hirsutum* cultivar difenmian168	MK792840	Yangtze river	S64
12	*Gossypium hirsutum* cultivar ekangmian7	MK792841	Yangtze river	S273
13	*Gossypium hirsutum* cultivar emian12(4947)	MK792842	Yangtze river	S263
14	*Gossypium hirsutum* cultivar gaochanbukangchong RRM	MK792843	Yangtze river	S246
15	*Gossypium hirsutum* cultivar guangyedaizimian	MK792844	United States	S59
16	*Gossypium hirsutum* cultivar guokang12 (GK12)	MK792845	Yellow river	S156
17	*Gossypium hirsutum* cultivar hanmian802	MK792846	Yellow river	S162
18	*Gossypium hirsutum* cultivar humian204	MK792847	Yangtze river	S257
19	*Gossypium hirsutum* cultivar Jan-86	MK792848	Yellow river	S211
20	*Gossypium hirsutum* cultivar liaomian10	MK792849	North China	S234
21	*Gossypium hirsutum* cultivar lumianyan21(lu1138)	MK792850	Yellow river	S163
22	*Gossypium hirsutum* cultivar shan401	MK792851	Yellow river	S10
23	*Gossypium hirsutum* cultivar simian4	MK,792852	Yangtze river	S272
24	*Gossypium hirsutum* cultivar sizimian4	MK792853	United States	S38
25	*Gossypium hirsutum* cultivar sumian5	MK792854	Yangtze river	S45
26	*Gossypium hirsutum* cultivar xinluzhong7	MK792855	Northwest China	S275
27	*Gossypium hirsutum* cultivar xinluzhong9 (1318136-160)	MK792856	Northwest China	S277
28	*Gossypium hirsutum* cultivar xinluzhong10	MK792857	Northwest China	S278
29	*Gossypium hirsutum* cultivar xinluzhong19	MK792858	Northwest China	S281
30	*Gossypium hirsutum* cultivar xuzhou209	MK792859	Yangtze river	S13
31	*Gossypium hirsutum* cultivar yanmian48	MK792860	Yangtze river	S265
32	*Gossypium hirsutum* cultivar youLU272⊕	MK792861	Yellow river	S175
33	*Gossypium hirsutum* cultivar yumian1	MK792862	Yangtze river	S271
34	*Gossypium hirsutum* cultivar zhong053	MK792863	Yangtze river	S8
35	*Gossypium hirsutum* cultivar zhongzhimian GD89	MK792864	Yellow river	S185
36	*Gossypium barbadense* cultivar zhonghai7	HQ901199	NCBI	AD_{2_99}
37	*Gossypium barbadense* cultivar kaiyuan	HQ901200	NCBI	AD_{2_200}
38	*Gossypium barbadense* cultivar yuanmou	HQ901198	NCBI	AD_{2_98}
39	*Gossypium darwinii*	NC_016670	NCBI	AD_{5_70}
40	*Gossypium tomentosum*	NC_016690	NCBI	AD_{3_90}
41	*Gossypium mustelinum*	NC_016711	NCBI	AD_4
42	*Gossypium hirsutum*	NC_007944	NCBI	AD_{1_44}
43	*Gossypium barbadense*	NC_008641	NCBI	AD_{2_41}
44	*Gossypium africanum*	NC_016692	NCBI	A_{1_a}
45	*Gossypium arboreum*	NC_016712	NCBI	A_2
46	*Gossypium longicalyx*	JF317354	NCBI	F_1
47	*Gossypium anomalum*	JF317356	NCBI	B_1
48	*Gossypium capitis-viridis*	NC_018111	NCBI	B_3
49	*Gossypium sturtianum*	JF317353	NCBI	C_1
50	*Gossypium nandewarense*	MG779276	Sanya, Hainan, China	C_{1-n}
51	*Gossypium robinsonii*	NC_018113	NCBI	C_2
52	*Gossypium bickii*	JF317352	NCBI	G_1
53	*Gossypium australe*	NC_033401	NCBI	G_2

(Continued)

TABLE 1 | (Continued)

Number	Species	Accession number	Source	Logogram
54	*Gossypium populifolium*	NC_033398	NCBI	K_2
55	*Gossypium thurberi*	JF317353	NCBI	D_1
56	*Gossypium armourianum*	MG891801	Sanya, Hainan, China	D_{2-1}
57	*Gossypium harknessii*	NC_033333	NCBI	D_{2-2}
58	*Gossypium klotzschianum*	NC_033394	NCBI	D_{3-k}
59	*Gossypium davidsonii*	NC_033395	NCBI	D_{3-d}
60	*Gossypium aridum*	NC_033396	NCBI	D_4
61	*Gossypium raimondii*	NC_016668	NCBI	D_5
62	*Gossypium gossypioides*	NC_017894	NCBI	D_6
63	*Gossypium lobatum*	MG891802	Sanya, Hainan, China	D_7
64	*Gossypium trilobum*	MG800783	Sanya, Hainan, China	D_8
65	*Gossypium laxum*	KF806549	NCBI	D_9
66	*Gossypium turneri*	NC_026835	NCBI	D_{10}
67	*Gossypium schwendimanii*	MG891803	Sanya, Hainan, China	D_{11}
68	*Gossypium stooksii*	JF317354	NCBI	E_1
69	*Gossypium somalense*	NC_018110	NCBI	E_2
70	*Gossypium areyiabum*	NC_018112	NCBI	E_3
71	*Gossypium incanum*	NC_018109	NCBI	E_4
72	*Gossypium latifolium*	MG800784	Sanya, Hainan, China	kym

The following parameters were employed: statistical method, maximum likelihood; analysis, substitution pattern estimation (MCL); substitution type, nucleotides; scope, all selected taxa; model/method, Tamura–Nei (automatic selection); gaps/missing data treatment, partial deletion, and site coverage cut off (%), 95 (Mohanta and Bae, 2017). Finally, we converted the transition/transversion rates for the two groups into two histograms. In addition, DnaSP v5.10 (Librado and Rozas, 2009) was used to calculate the non-synonymous (dN) and synonymous (dS) mutations in coding regions for the two groups.

Estimation of Mutation Rates

The two cotton groups described above were also used to calculate the mutation rates. The rate of mutation per site per year (μ) was estimated using the formula: $\mu = m/(nT)$, where m is the number of observed mutations, n is the number of total sites, and T is the divergence time of a node (Denver et al., 2009). The μ values for structural mutations were calculated using the method described by Saiton and Ueda (Saitou and Ueda, 1994), where the total number of structural mutations was divided by the additive time based on the branch lengths and by the length of the nucleotide sequences. Finally, we calculated the evolutionary rates for nucleotide substitutions and indels. The indel rates were calculated for the two groups using DnaSP v5.10 (Librado and Rozas, 2009).

Selection Pressure Analysis

To identify domestication selected genes, we performed selection pressure analysis using the Codeml program (Yang et al., 2005) and two different groups of genotypes, where one group comprised the wild diploid cotton species with a total of 28 genotypes and the other group contained the upland cotton semi-wild races and cultivated varieties with a total of 37 cotton genotypes (excluding *G. barbadense* and its three cultivated accessions, i.e., *G. tomentosum*, *G. mustelinum* and *G. darwinii*, because these seven genotypes were not involved in the domestication selection process for upland cotton). In general, the non-synonymous (dN) and synonymous substitution (dS) rate ratio (ω = dN/dS) was sensitive to selection pressure during evolution at the protein level, and it was particularly useful for identifying positive selection. Geneious v8.0.2 (Kearse et al., 2012) and MAFFT v7.0.0 (Katoh and Standley, 2013) were used to extract and align 77 protein-coding chloroplast genes from the two groups. Maximum likelihood phylogenetic trees were constructed based on the complete chloroplast genome sequences using RAxML v7.2.8 (Stamatakis, 2006). This model allowed the ω ratio to vary among sites with a fixed ω ratio for the whole tree to test for site-specific evolution in the gene phylogeny (Yang and Nielsen, 2002). Log-likelihood values of every model were compared against a neutral model based on likelihood ratio tests in order to determine statistically significant differences. Only the candidate sites for positive selection with significant support based on the posterior probability (p of ($\omega > 1$) $\geq$0.99; Bayes Empirical Bayes approach) identified by M2 and M8 were considered further.

Diversity and Genetic Structure Analysis

DnaSP v5.10 (Librado and Rozas, 2009) was used to analyze the genetic diversity parameters based on the complete chloroplast genome sequences of seven semi-wild races and 29 cultivated upland cotton genotypes. We also calculated the haplotype diversity (H_d) (Nei and Tajima, 1981), nucleotide diversity (π) (Nei and Li, 1979), and the number of haplotypes (H) with DnaSP v5.10 software.

We also analyzed the genetic structure patterns using the Bayesian Markov chain Monte Carlo clustering analysis method implemented in STRUCTURE 2.3.3 (Pritchard et al., 2000; Falush et al., 2003; Hubisz et al., 2009). The admixture model with correlated allele frequencies was implemented for each run without a prior placed on the population information (Hubisz et al., 2009). We conducted eight independent runs for each value from K = 1–10 to estimate the "true" number of clusters in 200,000 Markov chain Monte Carlo cycles following a burn-in step of 500,000 iterations. The most likely number of clusters was defined using log probabilities [Pr(X| K)] (Pritchard et al., 2000) and the $\triangle K$ method (Evanno et al., 2005) *via* the online website STRUCTURE HARVESTER (Earl and VonHoldt, 2012). Next, CLUMPP 1.1.2 and the Greedy algorithm were used to align multiple runs of STRUCTURE for the same K value (Jakobsson and Rosenberg, 2007). Finally, we applied DISTRUCT 1.1 (Rosenberg, 2004) to graphically visualize the individual probabilities of cluster membership.

Gene Flow

We calculated the historical gene flow in semi-wild races and cultivated upland genotypes using Migrate-n (Beerli, 2006). First, we generated five independent Markov chain Monte Carlo cycles, each with 5,000,000 generations. We then sampled every 100 steps under a constant variation model and discarded the first 1,000,000 records as a burn-in and the other settings were at their default values. After checking for data convergence, we estimated the mode and 95% HPD (Du et al., 2017). In addition, we applied BAYESASS v3.0 to detect contemporary gene flow in the two groups (Wilson and Rannala, 2003). In these calculations, the three parameters comprising the migration rates (ΔM), allele frequencies (ΔA), and inbreeding coefficients (ΔF) were used as references to ensure that the optimal acceptance rates for the three parameters fell within the range of 20–60%. After continuous calculations, the correlation values for the genetic components were finally determined as 0.03, 0.16, and 0.14, respectively. We then conducted the analyses based on 5^7 iterations after a burn-in of 5^6 iterations and set 1,000 as the sampling frequency. Ten separate runs were performed to minimize the convergence problem (Feng et al., 2016). The method proposed by Meirmans was used to obtain the results with the lowest deviance (Meirmans, 2014).

RESULTS

Evolutionary Relationships

The chloroplast genome sequences and concatenated protein-coding genes were used to reconstruct the maximum likelihood phylogenetic relationships for 72 *Gossypium* genotypes, and the cotton relationships generated from the data sets had the same topology, as shown in **Figure 1**. The six major genetic clades identified comprised the A + AD, F, E, D, B, and C + G + K genomic groups. Interestingly, all of the cultivated upland cotton genotypes clustered with the semi-wild race latifolium, which also formed a large evolutionary lineage with the other semi-wild races. The A-genome cotton species and *G. barbadense* genotypes also formed a single clade and they were closest to the upland cotton branch, whereas the 13 D-genome species formed a strong monophyletic lineage. The Australian species (C + G + K) clustered into a small branch, which clustered into a large branch with the B-genome species. Four species representing the E-genomic group also clustered into a large evolutionary branch. These results were in good agreement with the biogeographic distributions of cotton species from different continents.

Divergence Time Estimation

The molecular dating showed that the divergence time between the genus *Gossypium* and outgroups (*B. ceiba* and *T. cacao)* was about 58.15 Mya (95% HPD = 56.53–60.04 Mya), which are consistent with previous estimates (Carvalho et al., 2011; **Figure 2**). The genus *Gossypium* originated about 11 Mya (95% HPD = 9.34–11.74 Mya) and most genomic groups in the genus diverged radially in a relatively narrow time range. Interestingly, the divergence time between the B-genome (African origin) and Australian clades (C + G + K) was estimated at 7.7 Mya (95% HPD = 6.3–9.8 Mya), which again supported the genetic relationship present in the B-genome, i.e., the B-genome branch and Australian branch were strongly grouped phylogenetically. The semi-wild races and cultivated upland cotton accessions diverged about 3.12 Mya and the ancestor of the D-genome originated at 5 Mya (95% HPD = 3.59–5.44 Mya). The divergence time of the allotetraploid AD clade was about 3.37 Mya (95% HPD = 2.44–4.93 Mya).

Nucleotide Substitutions

The ratios of transition/transversion were high among the semi-wild races and cultivated upland cotton genotypes (1.41), but low among the genotypes of the wild cotton species (1.16) (**Table 2**). There were significant differences in the proportions of two transition mutations and four transversion mutations between the two groups (**Figure 3**). Among the four transversion mutations, the proportion of A-C + T-G mutations was similar to that of C-G + G-C mutations in the groups. In addition, few A-T + T-A and C-A + G-T mutations were found in all combinations.

The 4,074 biallelic SNPs were subdivided into coding, intron, and intergenic spacer regions, and sorted into two groups comprising wild cotton species, and semi-wild and cultivated upland cotton genotypes (**Table 3**). In wild cotton group, there were 3,753 SNPs in total: 1,375 in coding regions, 264 in intron regions, and 2,693 in intergenic spacer regions. The percentages of SNP to the total lengths were 1.72, 1.22, 2.95, respectively, manifesting the intergenic spacer region sequences were more variable than the intron regions. In the coding regions, there were 1,027 non-synonymous mutations and 347 synonymous mutations, and the dN/dS was about 2.96. In the semi-wild and cultivated cotton genotypes, the sequences of the intergenic spacers and intron regions were more variable than the coding regions. The dN/dS ratio (3.5) was larger for this group than the wild cotton species (56 non-synonymous mutations and 16 synonymous mutations).

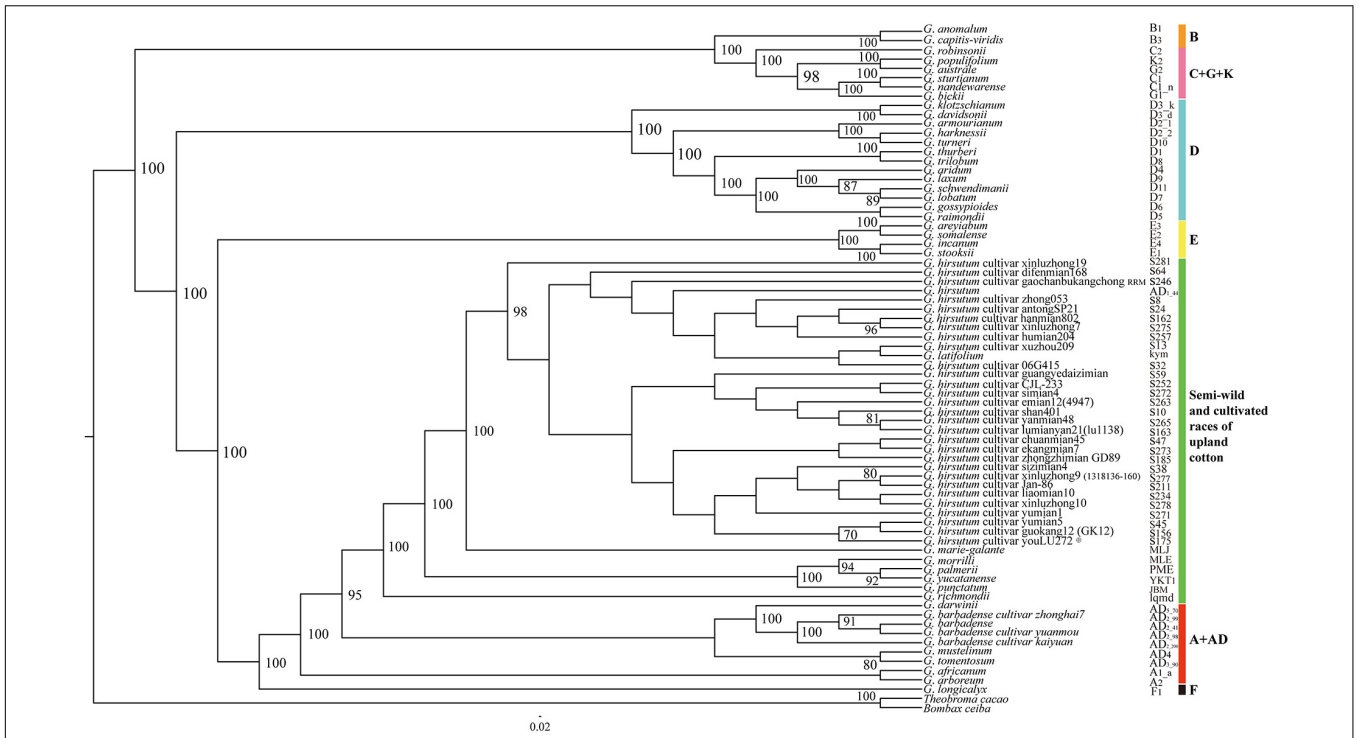

FIGURE 1 | Phylogenetic relationships among 72 Gossypium accessions based on complete chloroplast genomes. Green represents the cultivated accessions and semi-wild races of upland cotton, and other colors represent six genetic clades. *B. ceiba* and *T. cacao* were used as outgroups.

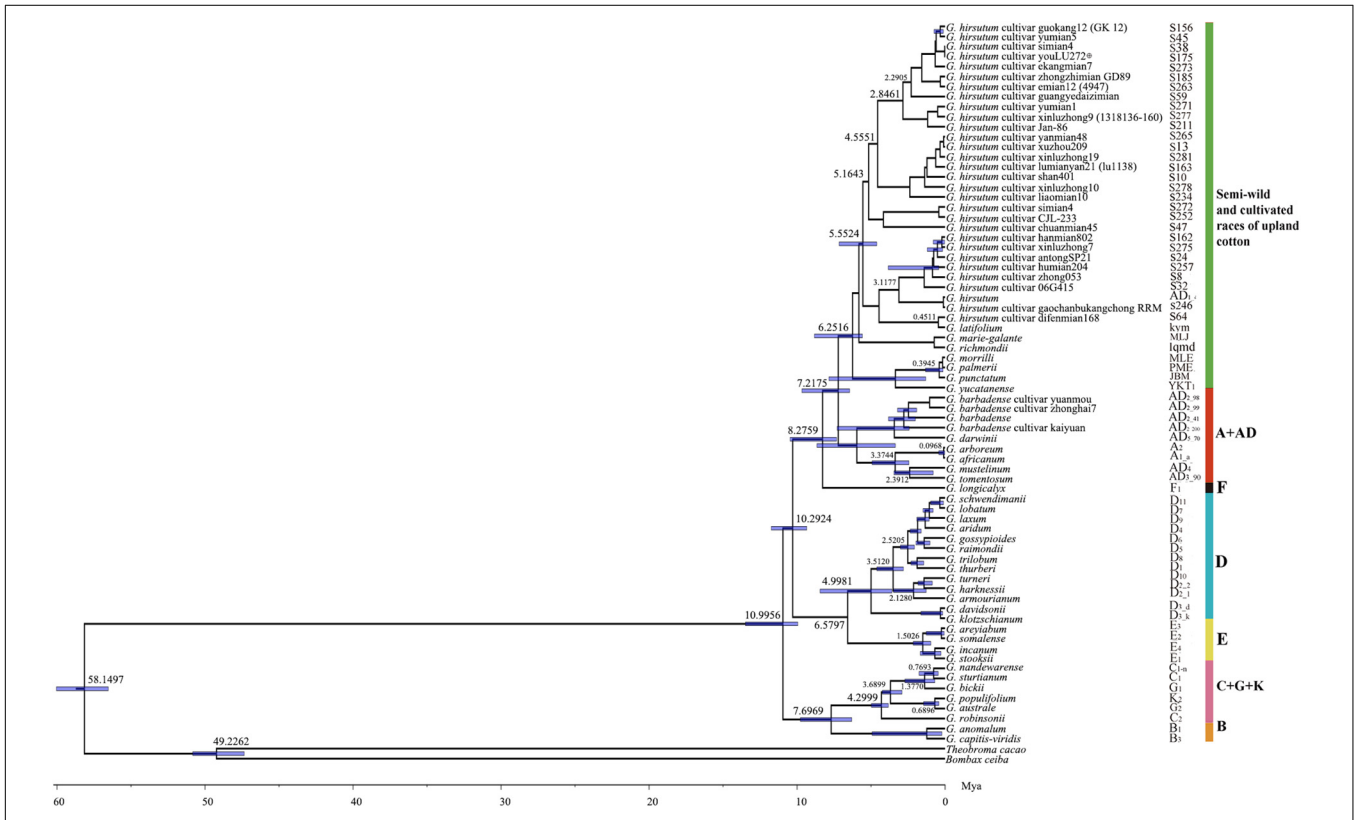

FIGURE 2 | Divergence time tree obtained for cotton accessions based on 72 chloroplast protein-coding sequences.

TABLE 2 | Ratios of transitions and transversions for plastid protein-coding sequences in cotton accessions.

	Semi-wild and cultivated cotton accessions					Wild cotton species				
From\To	A	T	C	G	Ts/Tv	A	T	C	G	Ts/Tv
A	–	4.2727	4.2655	**22.2241**	1.4100	–	4.6792	6.4654	**13.9443**	1.1600
T	3.3880	–	**13.9884**	8.5670		5.7307	–	**15.6745**	6.2325	
C	3.3880	**14.0120**	–	8.5670		5.7307	**11.3441**	–	6.2325	
G	**8.7889**	4.2727	4.2655	–		**12.8216**	4.6792	6.4654	–	

Rates of different transitional substitutions are shown in bold, whereas those of transversional substitutions are not shown in bold.

Estimation of Mutation Rate

The evolutionary rates were calculated based on the lengths of the genomes, number of substitutions, and times since divergence. In total, 1,375 substitutions were estimated in the wild species group and 77 in the semi-wild races and cultivated upland cotton group. The evolutionary rate of nucleotide substitutions was 1.2×10^{-9} per site per year in the wild species group compared with 0.18×10^{-9} per site per year in the semi-wild and cultivated group. In addition, 479 indels were identified in the wild cotton species and the evolutionary rate for indels was estimated at 0.4×10^{-9} per site per year. In the semi-wild and cultivated group, 24 indels were detected and the evolutionary rate was estimated at 0.05×10^{-11} per site per year.

Selection Pressures

We identified 16 genes with sites under positive selection in the wild species group (**Supplementary Tables 1, 2**). These genes comprised two ATP subunit genes (*atpB* and *atpE*), three ribosome small subunit genes (*rps2*, *rps3*, and *rps12*), three genes encoding cytochrome b/f complex subunit proteins (*petB*, *petD*, and *petN*), one NADH oxidoreductase gene (*ndhG*), one DNA-dependent RNA polymerase gene (*rpoC2*), one gene encoding ribosome large subunit protein (*rpl16*), and five other genes (*ccsA*, *cemA*, *rbcL*, *ycf1*, and *ycf2*). According to the M2 and M8 models, the *rps12* gene harbored 28 sites under positive selection, as well as 34 sites in *ycf2*, six and four sites in *ycf1*, two and five sites in *ndhG*, and one site each in the *ccsA*, *cemA*, *rpl16*, *rps3*, and *petB* genes. The M8 model detected 15 sites under positive selection in the *rps2* gene. However, sites under positive selection in the *atpB* (five), *atpE* (two), and *rbcL* (two) genes were only detected by the M2 model, and the other six genes had only one active site.

We only identified the ribosome large subunit protein (*rpl2*) gene with sites under positive selection in the semi-wild and cultivated group, where it harbored four sites under positive selection in the M2 model (**Supplementary Tables 3, 4**).

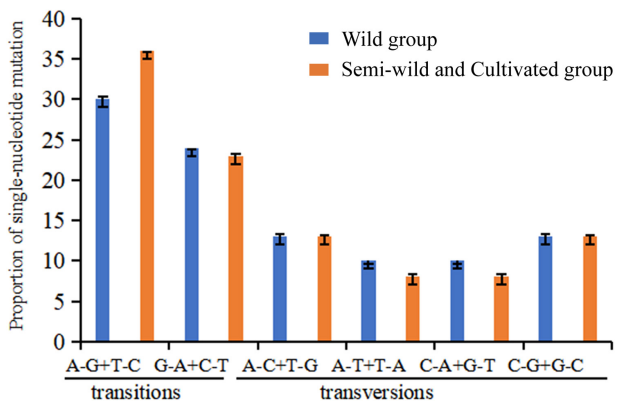

FIGURE 3 | Nucleotide substitution patterns in wild cotton species, semi-wild races, and cultivated cotton accessions based on SNP site variations. The patterns were divided into six types as indicated by the six non-strand-specific base substitution types. $p = 0.97681$. Because the calculated value is a fixed value with a decimal. We calculate the error bar between the actual value and the integer substitution site. The *p*-value is calculated by *T*-test.

Diversity and Genetic structure

Seven chloroplast DNA haplotypes were identified in the semi-wild races and 22 in the cultivated upland cotton genotypes (**Table 4**). The haplotypes diversity (H_d) and π values were slightly higher for the semi-wild races than the cultivated genotypes. STRUCTURE analyses and the ΔK statistic indicated an "optimal" value for K (number of populations modeled) of 2 (**Supplementary Figure 1**), thereby supporting the existence of two major clusters in the data set (**Figure 4**). The semi-wild races were primarily assigned to cluster I and the cultivated genotypes to cluster II, whereas the races marie-galante and latifolium had notable fractions assigned to cluster II, thereby suggesting genetic introgression between the two groups.

Gene Flow

Patterns of historical and contemporary gene flow were detected between the semi-wild and cultivated upland cotton genotypes. Migrate-n analysis showed that historical gene flow ranged from 149.77 (135.69–164.85) for the semi-wild group to 377.47 (344.25–413.03) for the cultivated group, thereby indicating asymmetric gene flow between the groups. Significant asymmetric contemporary gene flow was also found between the groups, where the values ranged from 0.1110 (0.0612–0.1608) for the semi-wild group to 0.0108 (0.0004–0.0212) for the cultivated group. These results suggest a higher level of historical gene flow during domestication compared with the low level of contemporary gene flow.

DISCUSSION

Evolutionary Relationships

Some previous studies have explored the molecular phylogenetic relationships of cotton, mostly based on a small number of

TABLE 3 | Taxonomic and genomic distribution of biallelic single nucleotide polymorphic loci in wild, semi-wild, and cultivated cotton plastid genomes.

		Wild accessions			Semi-wild and cultivated accessions	
Genome region	**Length (bp)**	**Value**	**%**	**Length (bp)**	**Value**	**%**
Total substitutions	163,400	3,753	2.3	166,237	321	0.19
Coding region	79,704	1,375	1.72	79,968	77	0.1
Non-synonymous	/	1,027	1.29	/	56	0.07
Synonymous	/	347	0.44	/	16	0.02
dN/dS	/	2.96	/	/	3.5	/
Intron	21,581	264	1.22	21,292	14	0.07
Intergenic spacer	81,381	2,693	2.95	77,524	130	0.17

TABLE 4 | Nucleotide diversity and haplotype frequencies for plastid genomes in semi-wild and cultivated accessions of upland cotton.

Population	**Number of samples**	**Number of haplotypes (H)**	**Hd (SD)**	**π (SD) × 100**	**Number of segregation sites**	**Theta**
Semi-wild races	7	7	1.000 (0.076)	0.00035 (0.00006)	157	0.196
Cultivated accessions	29	22	0.946 (0.035)	0.00010 (0.00003)	170	0.132

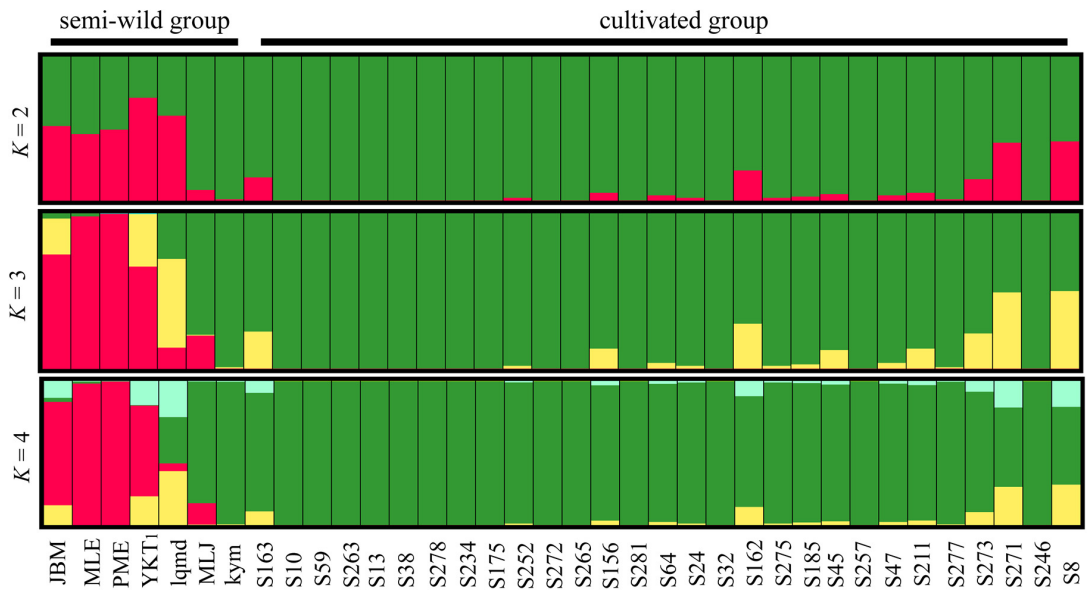

FIGURE 4 | Bayesian clustering results obtained for cotton accessions by STRUCTURE analysis. Each individual is represented by a vertical bar, which is partitioned into K colored segments showing the individual's probability of belonging to the group with that color.

plastids and nuclear DNA markers, as well as the complete chloroplast genome sequence and mitochondrial genome data set of a limited number of cotton species (Cronn et al., 2002; Senchina et al., 2003; Wendel et al., 2009; Xu et al., 2012; Wendel and Grover, 2015; Chen et al., 2016, 2017a,b; Wu et al., 2018). However, the relationship between cultivated accessions of upland cotton and other species of *Gossypium* is not clear now. Therefore, we built phylogenetic analyses on 72 cotton plastid genome sequences including wild species, semi-wild races and cultivated accessions of *Gossypium*, representing the largest number of known cotton species. In the phylogenetic tree, *Gossypium* species were primarily divided into three large genetic branches. The outer two branches mainly comprised diploid cotton species and the upland cotton clade formed the inner branch. One of the two outside branches included the Australian species with C, G, and K-genomes, American D-genome species, and African E- and B-genome species. Other studies have also shown that species with the G-genome have a common nested relationship with C-genome species, probably due to the frequent capture of chloroplasts in the *G. bickii* lineage (Seelanan et al., 1999; Liu et al., 2001). The other outside branch comprised the African F-genome species, Asian–African A-genome species, and American AD-genome wild species and cultivated *G. barbadense* genotypes. The large internal branch included all of the upland cotton cultivars and semi-wild races. The race latifolium clustered more closely with the upland cotton genotypes, which may suggest a classification error because the race yucatanense is considered the closest progenitor of cultivated

upland cotton. Some studies have reported that the maternal donor of the chloroplast genome for the allotetraploid species was the A-genome progenitor (Cronn et al., 2002; Chen et al., 2016, 2017a; Huang et al., 2020), and this was supported by our phylogenetic analysis. The latest research showed that the two A-genome species (*G. herbaceum* and *G. arboreum*) have evolved independently with no ancestor-progeny relationship (Huang et al., 2020). In addition, the phylogenetic tree showed that all 13 D-genome species clustered into a single lineage with high support and they were more distantly related to the upland cotton genotypes. Some D-genome species formed closely associated pairs, including *G. klotzschianum* (D_{3-k}) with *G. davidsonii* (D_{3-d}), *G. harknessii* (D_{2-2}) with *G. turneri* (D_{10}), *G. thurberi* (D_1) with *G. trilobum* (D_8), and *G. raimondii* (D_5) with *G. gossypioides* (D_6). These results are consistent with previous reports of phylogenetic relationships based on nuclear genetic markers and chloroplast genome sequences (Alvarez et al., 2005; Ulloa et al., 2013; Chen et al., 2017a; Wu et al., 2018; Huang et al., 2020). The difference in phylogenetic relationships may be caused by the different genetic characteristics of the DNA markers used.

Divergence Time Analysis

We estimated the divergence time of *Gossypium* species based on the plastid protein-coding sequences. The results showed that the diversification between *Gossypium* and *T. cacao* was found to have occurred about 58 Mya, which was consistent with previous inferred results (Wendel, 1989; Senchina et al., 2003; Carvalho et al., 2011; Chen et al., 2016). Interestingly, the divergence time was estimated at 7.7 Mya (95% HPD = 6.3–9.8 Mya) between the B-genome and Australian clade (C + G + K), which was similar to the rapid radiation time calculated for all other cotton branches after differentiation from Australian cotton species (Chen et al., 2016). In addition, the evolutionary time of the cotton ancestors was 11 Mya and cotton species then rapidly differentiated radially, where the differentiation time of most branches was 5–6 Mya. These results were largely consistent with those obtained in other molecular studies (Chen et al., 2016, 2017a,b). The differentiation time for the semi-wild races, cultivated upland cotton genotypes, and AD-genome was estimated at 6.25 Mya, and that estimated for the race latifoloum and *Gossypium hirsutum* cultivar difenmian168 was 0.45 Mya. We also found that the divergence time betweem semi-wild races and cultivated upland cotton accessions were about 3.12 Mya, thereby indicating that they may have differentiated recently. The evolutionary time for the allotetraploid upland cotton accessions was 6.25 Mya (6.4–9.7), which agrees with the results obtained in previous studies (Senchina et al., 2003; Wang et al., 2017; Ma et al., 2019; Huang et al., 2020), where it was domesticated at least 4,000 to 5,000 years ago and subsequently subjected to direct selection (Wang et al., 2017). To the best of our knowledge, the present study is the first to use the protein-coding sequences in the chloroplast genome to estimate the divergence dates of the whole *Gossypium* species including semi-wild races and cultivated upland cotton genotypes, although the results could be improved by larger phylogenetic analyses.

Genetic Mutation

Mutation is the ultimate source of genetic variation, the substrate of evolution (Nachman and Crowell, 2000; Zhang et al., 2020). A previous study suggested that the mutation/substitution rates varied between and within genomes (Mohanta and Bae, 2017), and that they were influenced by factors such as the nearest neighbor bases, chromosomal position, and the efficiency of the repair systems between the leading and lagging DNA strands. In general, the presence of similar bases or derivatives of similar bases facilitates the base replacement in the DNA repair process, and thus transitions occur more frequently than transversions (Mohanta and Bae, 2017). Our results of nucleotide sequence evolution analysis showed that the transition rate was higher than the transversion rate for the cotton genotypes evaluated, which is consistent with previous reports (Mohanta and Bae, 2017; Mohanta et al., 2019). SNP represents the most common form of polymorphism in biological genomes. Common polymorphisms are effective genetic markers related to biological evolution (Zhang et al., 2020). In the present study, we identified 4,074 SNPs in the *Gossypium* cp genomes. Among them, there were more SNPs in the intergenic region than the intron region, indicating that intergenic spacer sequences were more variable than intron regions in the plastid genome, which was consistent with the latest research results (Zhang et al., 2020). Furthermore, the dN/dS ratios were larger than 1, thereby indicating that non-synonymous mutations were fixed in the genomes, which may be due to component-driven mutation pressure (Foster et al., 1997). The dN/dS ratios were higher for the semi-wild and cultivated upland cotton genotypes than those determined for the wild cotton species, which may suggest that upland cotton has been subject to very strong artificial selection during domestication. The results of evolutionary rates indicated that the rates of nucleotide substitutions and indels were higher in wild species than the upland genotypes, thereby suggesting that the semi-wild and cultivated upland genotypes might have evolved more slowly after speciation. Due to the influence of artificial domestication, the cultivated genotypes exhibited less variation with fewer mutations. Previous studies have shown that selection can act on the mutation rate (Baer et al., 2007). Moreover, according to our results, the mutation rate was lower for indels than nucleotide substitutions, which is consistent with a previous report (Wu et al., 2018).

Domestication Selection

By the mid-18th century, the coastal colonies of the southeastern United States had developed upland and Sea Island cotton varieties, which showed a long history of cotton domestication and breeding (Du et al., 2017). Evidence suggested that the domestication and breeding of allotetraploid cotton were superior to A-genomic diploid cotton in yield and quality (Hovav et al., 2008). And the allopolyploid cultivated cotton was first domesticated about 5,000 years ago (Yoo et al., 2014). Generally, synonymous and non-synonymous nucleotide substitutions are important markers of gene evolution. In most genes, synonymous nucleotide substitutions have occurred more frequently than non-synonymous substitutions (Ogawa et al., 1999). The rates

of non-synonymous and synonymous substitutions are relatively slow in plant chloroplast genomes because of purifying and neutral selection (Erixon and Oxelman, 2008; Ivanova et al., 2017). In the present study, selection pressure analysis identified 16 genes with sites under positive selection in the wild species group, but only one of these genes (*rpl2*) was identified in the semi-wild and cultivated group. We conclude that the selection pressure on semi-wild and cultivated cotton species has fewer genes at positive selection sites, whereas the wild species retained adaptive genes and the selected sites increased. These results are generally consistent with those obtained in previous studies of the effects of artificial domestication on selection pressure (Price, 2002). When plants experience relatively large changes in the environment, such as artificial domestication or natural selection, the relaxation of selection for certain characteristics is inevitable (Coss, 1999; Price, 2002). Thereby, humans would expect that natural selection of these features would lose its strength (Price, 2002). The *rpl2* domestication selection gene identified in semi-wild and cultivated cotton species may have played an important role in the adaptation of *Gossypium* to various environments (Price, 2002; Fan et al., 2018; Wu et al., 2018; Chen et al., 2020). Moreover, selection pressure analysis for wild and domesticated cotton species can provide novel insights into how human selection has affected duplicated genes in allopolyploids (Yoo et al., 2014; Chen et al., 2020). It is known that many important crops such as potato, wheat and soybean are obvious polyploids, so studying the genes of allopolyploid cotton may provide new insights into the role of polyploids in crop evolution (Yoo et al., 2014).

Genetic Diversity

Additionally, genetic diversity is the basis of crop improvement (Akter et al., 2019). Therefore, understanding the genetic diversity, structure, and relationships between varieties of upland cotton is very important for breeding (Fang et al., 2013). The semi-wild races exhibited higher nucleotide diversity ($H_d = 1.000$, $\pi = 0.00035$) than the cultivated genotypes ($H_d = 0.946$, $\pi = 0.00010$), thereby suggesting that artificial domestication reduced the chloroplast genetic diversity, which is consistent with a previous report (Ma et al., 2019). The low level of genetic diversity determined in the cultivated upland cotton accessions was primarily due to several genetic bottlenecks during the domestication process (Fang et al., 2013; Wang et al., 2017). Various studies have also suggested that the genetic basis of cultivated upland cotton genotypes is narrow (Abdurakhmonov et al., 2008; Campbell et al., 2009; Akter et al., 2019), although the diversity of derived cultivars obtained by various breeding methods is still evident. In addition, cotton breeding often involves hybridization and re-selection with a small number of breeding materials, thereby resulting in a loss of genetic diversity (Tyagi et al., 2014). The genetic structure is mainly affected by geographical isolation and genetic exchange isolation (Guo et al., 1997; Gutierrez et al., 2002). Genetic structure analysis showed that the semi-wild races and cultivated upland accessions were divided into two groups when $K = 2$. We observed that the seven semi-wild races and cultivated upland accessions exhibited significant admixture, that was, the two semi-wild races Marie-galante and latifolium had notable fractions assigned to cultivated accessions group, which indicated that the race latifolium had closest relationships with cultivated accessions, followed by the race marie-galante race, thereby indicating the introgression of a certain gene between the semi-wild races and cultivated accessions, or possibly germplasm sharing (Tyagi et al., 2014). These results were consistent with a previous study on increasing human-mediated effects leading to significantly genetic introgression (Du et al., 2017). A previous study also showed that the existence of this mixture may be related to the domestication history and the frequent appearance of superior genotypes in different breeding programs (Mulugeta et al., 2018). China is not a natural cotton-growing region, and thus many cotton genotypes, such as Foster, STV, DPL, Trice, King, and Uganda, have been introduced as extensive genetic sources for upland cotton varieties in China from several overseas sources for improving varieties (Chen and Du, 2006; Du et al., 2007; Jia et al., 2014a; Mulugeta et al., 2018). It is important to study the diversity and genetic structure of upland cotton genotypes as well as their relationships to facilitate the conservation and improvement of cotton (Mulugeta et al., 2018). In addition, the genetic diversity and population structure of upland cotton germplasm resources can be effectively used for genetic breeding, and it is of great significance for the systematic utilization of long-term genetic variation of upland cotton (Tyagi et al., 2014).

Genetic Introgression

Ancient gene flow between domesticated varieties and their wild relatives probably occurred historically through seed transmission, and it was possibly influenced by human activities and environmental events (Wegier et al., 2011). In the present study, asymmetric historical gene flow was determined between the semi-wild and cultivated upland genotypes, which is consistent with a previous study (Deynze et al., 2011). However, contemporary gene flow was greatly reduced, which may have been due to current isolation. Genetic studies of species in the early stages of domestication have identified multiple domestication origins or high levels of sustained gene flow between wild and cultivated genotypes (Gross and Olsen, 2010). A previous study also suggested that the genetic structure of upland cotton genotypes was weak or an admixture, which may have resulted in a strong historical gene flow (Epps et al., 2013). In general, gene flow is an important factor that affects the population structure over time, where it may reduce local adaptation by homogenizing the populations found in different environments or by spreading harmful alleles between populations. Gene flow might also contribute to the introduction of potential adaptive alleles into populations and increased genetic variation (Sexton et al., 2011; Epps and Keyghobadi, 2015; Welt et al., 2015). Some studies have also indicated that gene flow from cultivated upland genotypes to wild cotton tetraploid species has increased the risk of extinction for these wild species (Wegier et al., 2011). I.e., some wild cotton species *G. tomentosum* (in Hawaii), *G. mustelinum* (in Brazil) and *G. darwinii* (in Galapagos) were in danger of extinction as a result of hybridization with domesticated tetraploid cotton (Ellstrand, 2003; Simard, 2010). In addition, numerous studies have shown

that interspecific hybrids (*G. hirsutum* x *G. barbadense*) can serve as genetic links for gene transfer from domesticated cotton to other wild relatives (*G. darwinii*) (Ellstrand, 2003; Simard, 2010). This occurred during or after speciation lead to the retention of ancestral polymorphism due to incomplete lineage sorting (Heckman et al., 2007; Wilyard et al., 2009), or introgression or introgressive hybridization of previously geographically isolated species resulting from the genetic exchange after secondary contact (Liston et al., 1999; Gay et al., 2007). Moreover, among the four cultivated *Gossypium* plants, upland cotton exhibits the highest level of gene flow (Wendel et al., 1992; Abdurakhmonov et al., 2008), which is related to the strong artificial domestication that it has undergone. The extensive gene flow and/or genetic introgression among cotton accessions might have provided the novel genetic resources of cotton breeding. Therefore, the suitable management and conservation of different cotton species accessions are important in the future.

CONCLUSION

In conclusion, our phylogenetic analysis confirms the evolutionary relationship within the whole *Gossypium*, especially the relationships between semi-wild races and cultivated accessions were well resolved. We also identified that the *rpl2* gene was positively selected in semi-wild races and cultivated genotypes. Meanwhile, we found that the cultivated genotypes have experienced very strong selection pressure. In addition, we found that the genetic diversity of cultivated accessions was low compared to wild ones due to artificial domestication. Through the analyses of genetic structure and gene flow, we concluded that there was a certain gene introgression between semi-wild races and cultivated accessions. The present research provided novel genetic resources for cotton breeding, as well as novel molecular mechanisms insights for the evolution and domestication of cotton species.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

TZ, NW, and Z-HL: data curation and writing – original draft. YW: formal analysis. TZ, NW, and X-FM: investigation. YW, X-LZ, B-GL, WL, J-JS, C-XW, and AZ: methodology. X-FM: resources and validation. TZ and NW: software. Z-HL: supervision and writing – review & editing. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by grants from the National Key R&D Program (2021YFF1000100), the Innovation Project of the Chinese Academy of Agricultural Sciences (CAAS-ASTIP-ICR-KP-2021-01), the Xinjiang Tianshan Talents Program (2021), the Central Public-interest Scientific Institution Basal Research Fund (Y2021XK12), the Project of Introduction High-level Talents in Xinjiang Uygur Autonomous Region Flexible Talents (2020), the Shaanxi Science and Technology Innovation Team (2019TD-012), and the Key Program of Research and Development of Shaanxi Province (2022ZDLSF06-02), and the Public Health Specialty in the Department of Traditional Chinese Medicine (2019-39).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.873788/full#supplementary-material

REFERENCES

Abdurakhmonov, I., Kohel, R., Yu, J., Pepper, A., Abdullaev, A., Kushanov, F., et al. (2008). Molecular diversity and association mapping of fiber quality traits in exotic *G. hirsutum* L. germplasm. *Genomics* 92, 478–487. doi: 10.1016/j.ygeno.2008.07.013

Akter, T., Islam, A. K. M. A., Rasul, M. G., Kundu, S., Khalequzzaman, J. U., and Ahmed, J. U. (2019). Evaluation of genetic diversity in short duration cotton (*Gossypium hirsutum* L.). *J. Cotton Res.* 2:1. doi: 10.1186/s42397-018-0018-6

Alvarez, I., Cronn, R., and Wendel, J. F. (2005). Phylogeny of the new world diploid cottons (*Gossypium* L., Malvaceae) based on sequences of three low-copy nuclear genes. *Plant Syst. Evol.* 252, 199–214. doi: 10.1007/s00606-004-0294-0

Baer, C. F., Miyamoto, M. M., and Denver, D. R. (2007). Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nat. Rev. Genet.* 8, 619–631. doi: 10.1038/nrg2158

Beerli, P. (2006). Comparison of bayesian and maximum likelihood inference of population genetic parameters. *Bioinformatics* 22, 341–345. doi: 10.1093/bioinformatics/bti803

Brubaker, C. L., and Wendel, J. F. (1994). Re-evaluating the origin of domesticated cotton (*Gossypium hirsutum*: Malvaceae) using nuclear restriction fragment length polymorphisms (RFLPs). *Am. J. Bot.* 81, 1309–1326. doi: 10.2307/2445407

Burger, J. C., Chapman, M. A., and Burke, J. M. (2008). Molecular insights into the evolution of crop plants. *Am. J. Bot.* 95, 113–122. doi: 10.2307/27733400

Campbell, B. T., Williams, V. E., and Park, W. (2009). Using molecular markers and field performance data to characterize the Pee Dee cotton germplasm resources. *Euphytica* 169, 285–301. doi: 10.1007/s10681-009-9917-4

Carvalho, M. R., Herrera, F. A., Jaramillo, C. A., Wing, S. L., and Callejas, R. (2011). Paleocene malvaceae from northern South America and their biogeographical implications. *Am. J. Bot.* 98, 1337–1355. doi: 10.3732/ajb.1000539

Chen, G., and Du, X. M. (2006). Genetic diversity of source germplasm of upland cotton in China as determined by SSR marker. *Acta Genet. Sinica* 33, 733–745. doi: 10.1016/S0379-4172(06)60106-6

Chen, S. L., Pang, X. H., Song, J. Y., Shi, L. C., Yao, H. W., Han, J. P., et al. (2014). A renaissance in herbal medicine identification: from morphology to DNA. *Biotechnol. Adv.* 32, 1237–1244. doi: 10.1016/j.biotechadv.2014.07.004

Chen, Z. W., Feng, K., Grover, C. E., Li, P., Liu, F., Wang, Y. M., et al. (2016). Chloroplast DNA structural variation, phylogeny, and age of divergence among diploid cotton species. *PLoS One* 11:e0157183. doi: 10.1371/journal.pone.0157183

Chen, Z., Grover, C. E., Li, P. B., Wang, Y. M., Nie, H. S., Zhao, Y. P., et al. (2017a). Molecular evolution of the plastid genome during diversification of the cotton genus. *Mol. Phylogenet. Evol.* 112, 268–276. doi: 10.1016/j.ympev.2017.04.014

Chen, Z., Nie, H., Grover, C. E., Wang, Y., Li, P., Wang, M., et al. (2017b). Entire nucleotide sequences of *Gossypium* raimondii and *G. arboreum* mitochondrial genome srevealed a-genome speciesas cytoplasmic donor of the allotetraploid species. *Plant Biol.* 19, 484–493. doi: 10.1111/plb.12536

Chen, Z. J., Sreedasyam, A., Ando, A., Song, Q., Santiago, L. M., Hulse-Kemp, A. M., et al. (2020). Genomic diversifications of five *Gossypium* allopolyploid species and their impact on cotton improvement. *Nat. Genet.* 52, 525–533. doi: 10.1038/s41588-020-0614-5

Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A. J., Müller, W. E., Wetter, T., et al. (2004). Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res.* 14, 1147–1159. doi: 10.1101/gr.1917404

Coss, R. G. (1999). "). Effects of relaxed natural selection on the evolution of behavior. geographic variation," in *Behavior: Perspectives on Evolutionary Mechanisms*, eds S. A. Foster and J. A. Endler (New York, NY: Oxford University Press). 180–208. doi: 10.1099/vir.0.81834-0

Cronn, R. C., Small, R. L., Haselkorn, T., and Wendel, J. F. (2002). Rapid diversification of the cotton genus (*Gossypium: Malvaceae*) revealed by analysis of sixteen nuclear and chloroplast genes. *Am. J. Bot.* 89, 707–725. doi: 10.3732/ajb.89.4.707

Denver, D. R., Dolan, P. C., Wilhelm, L. J., Sung, W., Lucas-Lledo, J. I., Howe, D. K., et al. (2009). A genome-wide view of Caenorhabditis elegans base-substitution mutation processes. *Proc. Natl. Acad. Sci. U.S.A.* 106, 16310–16314. doi: 10.1073/pnas.0904895106

Deynze, A. E. V., Hutmacher, R. B., and Bradford, K. J. (2011). Gene flow between *Gossypium hirsutum* L. and *Gossypium barbadense* L. is asymmetric. *Crop Sci.* 51, 298–305. doi: 10.2135/cropsci2010.04.0213

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., 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, X. M., Zhou, Z. L., Jia, Y. H., and Liu, G. Q. (2007). Collection and conservation of cotton germplasm in China (english abstract). *Cotton Sci.* 19, 346–353.

Du, F. K., Hou, M., Wang, W., Mao, K., and Hampe, A. (2017). Phylogeography of *Quercus aquifolioides* provides novel insights into the Neogene history of a major global hotspot of plant diversity in South-West China. *J. Biogeogr.* 44, 294–307. doi: 10.1111/jbi.12836

Du, X., Huang, G., He, S., Yang, Z., Sun, G., Ma, X., et al. (2018). Resequencing of 243 diploid cotton accessions based on an updated a genome identifies the genetic basis of key agronomic traits. *Nat. Genet.* 50, 796–802. doi: 10.1038/s41588-018-0116-x

Earl, D. A., and VonHoldt, B. M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the evanno method. *Conserv. Genet. Resour.* 4, 359–361. doi: 10.1007/s12686-011-9548-7

Ellstrand, N. C. (2003). Dangerous liaisons-when cultivated plants mate with their wild relatives. *Plant Sci.* 167, 187–188. doi: 10.1016/j.plantsci.2004.02.020

Epps, C. W., Castillo, J. A., Schmidt-Kuntzel, A., du Preez, P., Stuart-Hill, G., Jago, M., et al. (2013). Contrasting historical and recent gene flow among African buffalo herds in the Caprivi Strip of Namibia. *J. Hered.* 104, 172–181. doi: 10.1093/jhered/ess142

Epps, C. W., and Keyghobadi, N. (2015). Landscape genetics in a changing world: disentangling historical and contemporary influences and inferring change. *Mol. Ecol.* 24, 6021–6040. doi: 10.1111/mec.13454

Erixon, P., and Oxelman, B. (2008). Whole-gene positive selection, elevated synonymous substitu-tion rates, duplication, and indel evolution of the chloroplast clpP1 gene. *PLoS One* 3:e1386. doi: 10.1371/journal.pone.0001386

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* 14, 2611–2620. doi: 10.1111/j.1365294X.2005.02553.x

Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164, 1567–1587. doi: 10.3410/f.1015548.197423

Fan, W. B., Wu, Y., Yang, J., Shahzad, K., and Li, Z. H. (2018). Comparative chloroplast genomics of dipsacales species: insights into sequence variation, adaptive evolution, and phylogenetic relationships. *Front. Plant Sci.* 9:689. doi: 10.3389/fpls.2018.00689

Fang, D. D., Hinze, L. L., Percy, R. G., Li, P., and Thyssen, G. (2013). A microsatellite-based genome-wide analysis of genetic diversity and linkage disequilibrium in upland cotton (*Gossypium hirsutum* L.) cultivars from major cotton-growing countries. *Euphytica* 191, 391–401. doi: 10.1007/s10681-013-0886-2

Fang, L., Wang, Q., Hu, Y., Jia, Y., Che, 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

Feng, L., Zheng, Q. J., Qian, Z. Q., Yang, J., Zhang, Y. P., Li, Z. H., et al. (2016). Genetic structure and evolutionary history of three alpine sclerophyllous oaks in east himalaya-hengduan mountains and adjacent regions. *Front. Plant Sci.* 7:1688. doi: 10.3389/fpls.2016.01688

Foster, P. G., Jermiin, L. S., and Hickey, D. A. (1997). Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. *J. Mol. Evol.* 44, 282–288. doi: 10.1007/PL00006145

Fryxell, P. A. (1969). A classification of *Gossypium* L. (Malvaceae). *Taxon* 18, 585–591. doi: 10.2307/1218405

Fryxell, P. A. (1978). *The Natural History of the Cotton Tribe (Malvaceae* tribe, *Gossypieae*). College Station, TX: Texas A&M University Press.

Gay, L., Neubauer, G., Zagalska-Neubauer, M., Debain, C., Pons, J. M., David, P., et al. (2007). Molecular and morphological patterns of introgression between two large white-headed gull species in a zone of recent secondary contact. *Mol. Ecol.* 16, 3215–3227. doi: 10.1111/j.1365-294X.2007.03363.x

Gepts, P. (2004). Crop domestication as a dong-term selection experiment. *Plant Breed. Rev.* 24, 1–44. doi: 10.1002/9780470650288.ch1

Gross, B. L., and Olsen, K. M. (2010). Genetic perspectives on crop domestication. *Trends Plant Sci.* 15, 529–537. doi: 10.1016/j.tplants.2010.05.008

Grover, C. E., Kim, H. R., Wing, R. A., Paterson, A. H., and Wendel, J. F. (2007). Microcolinearity and genome evolution in the AdhA region of diploid and polyploid cotton (*Gossypium*). *Plant J.* 50, 995–1006. doi: 10.1111/j.1365-313X.2007.03102.x

Gover, C. E., Pan, M., Yuan, D., Arick, M. A., Hu, G., Brase, L., et al. (2020). The *Gossypium* longicalyx genome as a resource for cotton breeding and evolution. *G3: Genes Genom. Genet.* 10, 1457–1467. doi: 10.1534/g3.120.401050

Guo, W. Z., Zhang, T. Z., Pan, J. J., and Wang, X. Y. (1997). A preliminary study on genetic diversity of Upland cotton cultivars in China. *Acta Gossypii. Sinica.* 9, 19–24.

Gutierrez, O. A., Basu, S., Saha, S., Jenkins, J. N., Shoemaker, D. B., Cheatham, C. L., et al. (2002). Genetic distance among selected cotton genotypes and its relationship with F2 performance. *Crop Sci.* 42, 1841–1847. doi: 10.2135/cropsci2002.1841

Hahn, C., Bachmann, L., and Chevreux, B. (2013). Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads-a baiting and iterative mapping approach. *Nucleic Acids Res.* 41:e129. doi: 10.1093/nar/gkt371

Heckman, K. L., Mariani, C. L., Rasoloarison, R., and Yoder, A. D. (2007). Multiple nuclear loci reveal patterns of incomplete lineage sorting and complex species history within western mouse lemurs (*Microcebus*). *Mol. Phylogenet. Evol.* 43, 353–367. doi: 10.1016/j.ympev.2007.03.005

Hu, G., Koh, J., Yoo, M. J., Grupp, K., Chen, S., and Wendel, J. F. (2013). Proteomic profiling of developing cotton fibers from wild and domesticated *Gossypium barbadense*. *New Phytol.* 200, 570–582. doi: 10.1111/nph.12381

Huang, G., Wu, Z., Percy, R. G., Bai, M., Li, Y., Frelichowski, J. E., et al. (2020). Genome sequence of *Gossypium* herbaceum and genome updates of *Gossypium* arboreum and *Gossypium hirsutum* provide insights into cotton a-genome evolution. *Nat. Genet.* 52, 516–524. doi: 10.1038/s41588-020-0607-4

Hovav, R., Chaudhary, B., Udall, J. A., Flagel, L., and Wendel, J. F. (2008). Parallel domestication, convergent evolution and duplicated gene recruitment in allopolyploid cotton. *Genetics* 179, 1725–1733. doi: 10.1534/genetics.108.089656

Hubisz, M. J., Falush, D., Stephens, M., and Pritchard, J. K. (2009). Inferring weak population structure with the assistance of sample group information. *Mol. Ecol. Resour.* 9, 1322–1332. doi: 10.1111/j.1755-0998.2009.02591.x

Iqbal, M. J., Reddy, O. U. K., El-Zik, K. M., and Pepper, A. E. (2001). A genetic bottleneck in the 'evolution under domestication' of upland cotton *Gossypium*

hirsutum L. examined using DNA fingerprinting. *Theor. Appl. Genet.* 103, 547–554. doi: 10.1007/PL00002908

Ivanova, Z., Sablok, G., Daskalova, E., Zahmanova, G., Apostolova, E., Yahubyan, G., et al. (2017). Chloroplast genome analysis of resurrection tertiary relict *Haberlea rhodo-pensis* highlights genes important for desiccation stress response. *Front. Plant Sci.* 8:204. doi: 10.3389/fpls.2017.00204

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

Jansen, R. K., Raubeson, L. A., Boore, J. L., Chumley, T. W., Haberle, R. C., Wyman, S. K., et al. (2007). Methods for obtaining and analyzing whole chloroplast genome sequences. *Methods Enzymol.* 395, 348–384. doi: 10.1016/S0076-6879(05)95020-9

Jia, Y. H., Sun, J. L., and Du, X. M. (2014a). "Cotton germplasm resources in China," in *World Cotton Germplasm Resources*, ed. I. Y. Abdurakhmonov (Uzbekistan: Academy of Sciences of Uzbekistan).

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

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

Li, F. G., Fan, G. Y., Wang, K. B., Sun, F. M., Yuan, Y. Y., and Song, G. L. (2014). Genome sequence of the cultivated cotton *Gossypium arboretum*. *Nat. Genet.* 46, 567–572. doi: 10.1038/ng.2987

Li, F. G., Fan, G. Y., Lu, C. R., Xiao, G. H., Zou, C. S., and Kohel, R. J. (2015). Genome sequence of cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. *Nat. Biotechnol.* 33, 524–530. doi: 10.1038/nbt.3208

Li, L., Hu, Y., He, M., Zhang, B., Wu, W., Cai, P., et al. (2021). Comparative chloroplast genomes: insights into the evolution of the chloroplast genome of *Camellia sinensis* and the phylogeny of *Camellia*. *BMC Genet.* 22:138. doi: 10.1186/s12864-021-07427-2

Li, Y. Z., Liu, Z., Zhang, K. Y., Chen, S. Y., and Zhang, Q. D. (2020). Genome-wide analysis and comparison of the DNA-binding one zinc finger gene family in diploid and tetraploid cotton (*Gossypium*). *PLoS One* 15:e0235317. doi: 10.1371/journal.pone.0235317

Librado, P., and Rozas, J. (2009). DnaSPv5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25, 1451–1452. doi: 10.1093/bioinformatics/btp187

Liu, Q., Brubaker, C. L., Green, A. G., Marshall, D. R., Sharp, P. J., and Singh, S. P. (2001). Evolution of the FAD2-1 fatty acid desaturase 5'UTR intron and the molecular systematics of *Gossypium* (*Malvaceae*). *Am. J. Bot.* 88, 92–102. doi: 10.2307/2657130

Liston, A., Robinson, W. A., Piñero, D., and Alvarez-Buylla, E. R. (1999). Phylogenetics of *Pinus (Pinaceae)* based on nuclear ribosomal DNA internal transcribed spacer region sequences. *Mol. Phylogenet. Evol.* 11, 95–109. doi: 10.1006/mpev.1998.0550

Ma, X. F., Wang, Z. Y., Li, W., Zhang, Y. Z., Zhou, X. J., Liu, Y. A., et al. (2019). Resequencing core accessions of a pedigree identifies derivation of genomic segments and key agronomic trait loci during cotton improvement. *Plant Biotechnol. J.* 17, 762–775. doi: 10.1111/pbi.13013

Mabry, M. E., Turner-Hissong, S. D., Gallagher, E. Y., McAlvay, A. C., An, H., Edger, P. P., et al. (2021). The Evolutionary History of Wild, Domesticated, and Feral *Brassica oleracea* (*Brassicaceae*). *Mol. Biol. Evol.* 38, 4419–4434. doi: 10.1093/molbev/msab183

May, O. L., Bowman, D. T., and Calhoun, D. S. (1995). Genetic diversity of US upland cotton cultivars released between 1980 and 1990. *Crop Sci.* 35, 1570–1574. doi: 10.2135/cropsci1995.0011183X003500060009x

Meirmans, P. G. (2014). Nonconvergence in bayesian estimation of migration rates. *Mol. Ecol. Resour.* 14, 726–733. doi: 10.1111/1755-0998.12216

Mohanta, T. K., and Bae, H. (2017). Analyses of genomic tRNA reveal presence of novel tRNAs in Oryza sativa. *Front. Genet.* 8:90. doi: 10.3389/fgene.2017.00090

Mohanta, T. K., Khan, A. L., Hashem, A., Allah, E. F. A., Yadav, D., and Al-Harrasi, A. (2019). Genomic and evolutionary aspects of chloroplast tRNA in monocot plants. *BMC Plant Biol.* 19:39. doi: 10.1186/s12870-018-1625-6

Mulugeta, S., Ming, D. X., Pu, H. S., Hua, J. Y., Zhaoe, P., and Ling, S. J. (2018). Analysis of genetic diversity and population structure in upland cotton (*Gossypium hirsutum* L.) germplasm using simple sequence repeats. *J. Genet.* 97, 513–522. doi: 10.1007/s12041-018-0943-7

Nachman, M. W., and Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics* 156, 297–304. doi: 10.1093/genetics/156.1.297

Nei, M., and Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. U.S.A.* 76, 5269–5273. doi: 10.1073/pnas.76.10.5269

Nei, M., and Tajima, F. (1981). DNA polymorphism detectable by restriction endonucleases. *Genetics* 97, 145–163. doi: 10.1007/BF00135050

Niu, E., Jiang, C., Wang, W., Zhang, Y., and Zhu, S. (2020). Chloroplast genome variation and evolutionary analysis of *Olea europaea L. Genes* 11:879. doi: 10.3390/genes11080879

Ogawa, T., Ishii, C., Kagawa, D., Muramoto, K., and Kamiya, H. (1999). Accelerated evolution in the protein-coding region of galectin cDNAs, congerin I and congerin II, from skin mucus of conger eel (*Conger myriaster*). *Biosci. Biotechnol. Biochem.* 63, 1203–1208. doi: 10.1271/bbb.63.1203

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

Patel, R. K., and Jain, M. (2012). NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PloS one.* 7:e30619. doi: 10.1371/journal.pone.0030619

Pfeil, B. E., and Crisp, M. D. (2008). The age and biogeography of citrus and the orange subfamily (*Rutaceae: Aurantioideae*) in Australasia and New Caledonia. *Am. J. Bot.* 95, 1621–1631. doi: 10.2307/41923047

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

Price, E. O. (2002). *Relaxation of Natural Selection*. California, CA: California Univ. Press.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959. doi: 10.1093/genetics/155.2.945

Rosenberg, N. A. (2004). Distruct: a program for the graphical display of population structure. *Mol. Ecol. Resour.* 4, 137–138. doi: 10.1046/j.1471-8286.2003.00566.x

Ruan, Y. L. (2003). Suppression of sucrose synthase gene expression represses cotton fiber cell initiation, elongation, and seed development. *Plant Cell* 15, 952–964. doi: 10.1105/tpc.010108

Saitou, N., and Ueda, S. (1994). Evolutionary rates of insertion and deletion in noncoding nucleotide sequences of primates. *Mol. Biol. Evol.* 11, 504–512. doi: 10.1016/0303-7207(94)90253-4

Schattner, P., Brooks, A. N., and Lowe, T. M. (2005). The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.* 33, 686–689. doi: 10.1093/nar/gki366

Seelanan, T., Brubaker, C. L., Stewart, J. M., Craven, L. A., and Wendel, J. F. (1999). Molecular systematics of Australian *Gossypium* section Grandicalyx (Malvaceae). *Syst. Bot.* 24:183. doi: 10.2307/2419548

Senchina, D. S., Alvarez, I., Cronn, R. C., Liu, B., Rong, J., Noyes, R. D., et al. (2003). Rate variation among nuclear genes and the age of polyploidy in *Gossypium*. *Mol. Biol. Evol.* 20, 633–643. doi: 10.1093/molbev/msg065

Sexton, J. P., Strauss, S. Y., and Rice, K. J. (2011). Gene flow increases fitness at the warm edge of a species' range. *Proc. Natl. Acad. Sci. U.S.A* 108, 11704–11709. doi: 10.1073/pnas.1100404108

Simard, M. J. (2010). Gene flow between crops and their wild relatives. *Evol. Appl.* 3, 402–403. doi: 10.1111/j.1752-4571.2010.00138.x

Simon, R. B., Justin, T. P., Joshua, A. U., William, S. S., Daniel, G. P., Mark, A. A., et al. (2016). Independent domestication of two old world cotton species. *Genome Biol. Evol.* 8, 1940–1947. doi: 10.1093/gbe/evw129

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

Tyagi, P., Gore, M., Bowman, D., Campbell, B., Udall, J., and Kuraparthy, V. (2014). Genetic diversity and population structure in the US upsland cotton (*Gossypium*

hirsutum L.). *Theor. Appl. Genet.* 127, 283–295. doi: 10.1007/s00122-013-2217-3

Ulloa, M., Abdurakhmonov, I. Y., Perez-m, C., Percy, R., and Stewart, J. M. (2013). Genetic diversity and population structure of cotton (*Gossypium* spp.) of the new world assessed by SSR markers. *Botany* 91, 251–259. doi: 10.1139/cjb-2012-0192

Wang, M., Tu, L., Lin, M., Lin, Z., Wang, P., Yang, Q., et al. (2017). Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication. *Nat. Genet.* 49, 579–587. doi: 10.1038/ng.3807

Wang, S., Shi, C., and Gao, L. Z. (2013). Plastid genome sequence of a wild woody oil species, *Prinsepia utilis*, provides insights into evolutionary and mutational patterns of Rosaceae chloroplast genomes. *PLoS One* 8:e73946. doi: 10.1371/journal.pone.0073946

Wang, M., Tu, L., Yuan, D., Zhu, D., Shen, C., Li, J., et al. (2019). Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and *Gossypium barbadense*. *Nat. Genet.* 51, 224–229. doi: 10.1038/s41588-018-0282-x

Wang, X., Zhang, Y., Wang, L., Pan, Z., and Du, X. (2020). Casparian strip membrane domain proteins in *Gossypium* arboreum: genome-wide identification and negative regulation of lateral root growth. *BMC Genom.* 21:340. doi: 10.1186/s12864-020-6723-9

Wegier, A., Pineyro-nelson, A., Alarcon, J., Galvez-mariscal, A., Alvarezbuylla, E. R., and Pinero, D. (2011). Recent long-distance transgene flow into wild populations conforms to historical patterns of gene flow in cotton (*Gossypium hirsutum*) at its centre of origin. *Mol. Ecol.* 20, 4182–4194. doi: 10.1111/j.1365-294X.2011.05258.x

Welt, R. S., Litt, A., and Franks, S. J. (2015). Analysis of population genetic structure and geneflow in an annual plant before and after a rapid evolutionary response to drought. *AoB Plants* 7:lv026. doi: 10.1093/aobpla/plv02

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

Wendel, J. F., Brubaker, C., Alvarez, I., Cronn, R., and Stewart, J. M. (2009). "Evolution and natural history of the cotton genus," in *Genetics and Genomics of Cotton*, ed. A. HPaterson (London: Springer Science). 3–22. doi: 10.1007/978-0-387-70810-2_1

Wendel, J. F., Brubaker, C. L., and Seelanan, T. (2010). "The origin and evolution of *Gossypium*," in . *Physiology of Cotton*. (Dordrecht: Springer Netherlands). 1–18. doi: 10.1007/978-90-481-3195-2_1

Wendel, J. F., and Cronn, R. C. (2003). Polyploidy and the evolutionary history of cotton. *Adv. Agron.* 78, 139–186. doi: 10.1016/s0065-2113(02)78004-8

Wendel, J. F., and Grover, C. E. (2015). "Taxonomy and evolution of the cotton genus, *Gossypium*," in *Cotton*, eds D. D. Fang and R. G. Percy (Madison, WI: American Society of Agronomy Inc.). 25–44. doi: 10.2134/agronmonogr57.2013.0020

Wendel, J. F., Brubaker, C. L., and Percival, A. E. (1992). Genetic diversity in *Gossypium hirsutum* and the origin of upland cotton. *Am. J. Bot.* 79, 1291–1310. doi: 10.1002/j.1537-2197.1992.tb13734.x

Wilson, G. A., and Rannala, B. (2003). Bayesian inference of recent migration rates using multilocus genotypes. *Genetics* 163, 1177–1191. doi: 10.1093/eurpub/13.1.11

Wilyard, A., Cronn, R., and Liston, A. (2009). Reticulate evolution and incomplete lineage sorting among the Ponderosa pines. *Mol. Phylogenet. Evol.* 52, 498–511. doi: 10.1016/j.ympev.2009.02.011

Wu, Y., Liu, F., Yang, D. G., Li, W., Zhou, X. J., Pei, X. Y., et al. (2018). Comparative chloroplast genomics of *Gossypium* species: insights into repeat sequence variations and phylogeny. *Front. Plant Sci.* 9:376. doi: 10.3389/fpls.2018.00376

Wyman, S. K., Jansen, R. K., and Boore, J. L. (2004). Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* 20, 3252–3255. doi: 10.1093/bioinformatics/bth352

Xu, Q., Xiong, G. J., Li, P. B., He, F., Huang, Y., Wang, K. B., et al. (2012). Analysis of complete nucleotide sequences of 12 *Gossypium* chloroplast genomes: origin and evolution of allotetraploids. *PLoS One* 7:e37128. doi: 10.1371/journal.pone.0037128

Yang, Z. H., and Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* 19, 908–917. doi: 10.1093/oxfordjournals.molbev.a004148

Yang, Z., Qanmber, G., Wang, Z., Yang, Z., and Li, F. (2020). *Gossypium* genomics: trends, scope, and utilization for cotton improvement. *Trends Plant Sci.* 25, 488–500. doi: 10.1016/j.tplants.2019.12.011

Yang, Z. H., Wong, W. S., and Nielsen, R. (2005). Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.* 22, 1107–1118. doi: 10.1093/molbev/msi097

Yoo, M. J., Wendel, J. F., and Bomblies, K. (2014). Comparative evolutionary and developmental dynamics of the cotton (*Gossypium hirsutum*) fiber transcriptome. *PLoS Genet.* 10:e1004073. doi: 10.1371/journal.pgen.1004073

Zhang, T. Z., Hu, Y., Jiang, W. K., Fang, L., Guan, X. Y., and Chen, J. D. (2015). Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. *Nat. Biotechnol.* 33, 531–537. doi: 10.1038/nbt.3207

Zhang, T. T., Zhang, N. Y., Li, W., Zhou, X. J., Pei, X. Y., Liu, Y. G., et al. (2020). Genetic structure, gene flow pattern, and association analysis of superior germplasm resources in domesticated upland cotton (*Gossypium hirsutum* L.). *Plant Diver.* 42, 189–197. doi: 10.1016/j.pld.2020.03.001

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

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

Copyright © 2022 Zhou, Wang, Wang, Zhang, Li, Li, Su, Wang, Zhang, Ma and Li. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 18 April 2022
doi: 10.3389/fpls.2022.853861

Genomic Analysis Based on Chromosome-Level Genome Assembly Reveals an Expansion of Terpene Biosynthesis of *Azadirachta indica*

Yuhui Du[1†], Wei Song[1†], Zhiqiu Yin[2†], Shengbo Wu[3], Jiaheng Liu[3], Ning Wang[1], Hua Jin[1]*, Jianjun Qiao[3,4]* and Yi-Xin Huo[1,5]*

[1] Key Laboratory of Molecular Medicine and Biotherapy, School of Life Sciences, Beijing Institute of Technology, Beijing, China, [2] National Engineering Laboratory for Efficient Utilization of Soil and Fertilizer Resources, College of Resources and Environment, Shandong Agricultural University, Tai'an, China, [3] Key Laboratory of Systems Bioengineering (Ministry of Education), School of Chemical Engineering and Technology, Tianjin University, Tianjin, China, [4] SynBio Research Platform, Collaborative Innovation Centre of Chemical Science and Engineering (Tianjin), Tianjin University, Tianjin, China, [5] Tobacco Research Institute, Chinese Academy of Agricultural Sciences, Qingdao, China

Azadirachta indica (neem), an evergreen tree of the Meliaceae family, is a source of the potent biopesticide azadirachtin. The lack of a chromosome-level assembly impedes an in-depth understanding of its genome architecture and the comparative genomic analysis of *A. indica*. Here, a high-quality genome assembly of *A. indica* was constructed using a combination of data from Illumina, PacBio, and Hi-C technology, which is the first chromosome-scale genome assembly of *A. indica*. Based on the length of our assembly, the genome size of *A. indica* is estimated to be 281 Mb anchored to 14 chromosomes (contig N50 = 6 Mb and scaffold N50 = 19 Mb). The genome assembly contained 115 Mb repetitive elements and 25,767 protein-coding genes. Evolutional analysis revealed that *A. indica* didn't experience any whole-genome duplication (WGD) event after the core eudicot γ event, but some genes and genome segment might likely experienced recent duplications. The secondary metabolite clusters, TPS genes, and CYP genes were also identified. Comparative genomic analysis revealed that most of the *A. indica*-specific TPS genes and CYP genes were located on the terpene-related clusters on chromosome 13. It is suggested that chromosome 13 may play an important role in the specific terpene biosynthesis of *A. indica*. The gene duplication events may be responsible for the terpene biosynthesis expansion in *A. indica*. The genomic dataset and genomic analysis created for *A. indica* will shed light on terpene biosynthesis in *A. indica* and facilitate comparative genomic research of the family Meliaceae.

Keywords: *Azadirachta indica*, chromosome-level assembly, comparative genomics, terpene biosynthesis, genome evolution

OPEN ACCESS

Edited by:
Jeremy Coate,
Reed College, United States

Reviewed by:
Sunil Kumar Sahu,
Beijing Genomics Institute (BGI), China
Liangsheng Zhang,
Zhejiang University, China

***Correspondence:**
Hua Jin
huajin@bit.edu.cn
Jianjun Qiao
jianjunq@tju.edu.cn
Yi-Xin Huo
huoyixin@bit.edu.cn

[†] 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 January 2022*
Accepted: *07 March 2022*
Published: *18 April 2022*

Citation:
Du Y, Song W, Yin Z, Wu S, Liu J, Wang N, Jin H, Qiao J and Huo Y-X (2022) Genomic Analysis Based on Chromosome-Level Genome Assembly Reveals an Expansion of Terpene Biosynthesis of Azadirachta indica. Front. Plant Sci. 13:853861. doi: 10.3389/fpls.2022.853861

INTRODUCTION

Azadirachta indica (neem) is a member of Meliaceae family, which is extensively studied for its bioactive products (Schmutterer, 1995). It grows natively on the Indian subcontinent and also in other countries such as Egypt and the Kingdom of Saudi Arabia. *A. indica* is a source of abundant limonoids and simple terpenoids which are responsible for its biological activity (Dai et al., 2001). Azadirachtin, the most important active compound in the neem tree, has been intensively studied because of its wide range of insecticidal properties and low toxicity in mammals (Ley, 1994). Additionally, the neem tree extracts also exhibit many pharmaceutical functions, such as anti-inflammatory, anticancer, antimicrobial, and antidiabetic activities (Soares et al., 2014; Abdelhady et al., 2015). A lot of studies have focused on the synthesis of azadirachtin, including chemosynthesis, hairy root culture, cell line culture, and callus culture (Veitch et al., 2007; Srivastava and Srivastava, 2013; Mithilesh and Rakhi, 2014; Rodrigues et al., 2014). However, these methods are either of low extraction efficiency or not environmentally friendly. Therefore, the reconstruction of biosynthetic pathway of azadirachtin for heterologous production is an alternative method.

An omics strategy is an effective method to study the biosynthesis of secondary metabolites. Transcriptomes of *A. indica* tissues (stem, leaf, flower, root, and fruit) have been sequenced, which paved the way to the potential synthetic pathway of azadirachtin and gene expression profiles in various organs. Draft genomes have also been sequenced, which led to a basic understanding of the genetic characteristics of *A. indica* (Krishnan et al., 2012, 2016; Kuravadi et al., 2015). However, the lack of a chromosome-level genome sequence has hindered a full understanding of the secondary metabolite biosynthesis and the evolution of *A. indica*. In addition, Meliaceae are known to produce around 1,500 structurally diverse limonoids, which have agricultural and medical values (Hodgson et al., 2019). A chromosome-level genome is essential for genome-wide studies of the Meliaceae family.

In this study, the first chromosome-level genome of *A. indica* was assembled through a combination of Illumina, PacBio, and Hi-C technology. Based on the assembled genome sequence and annotation, we characterized the history of gene and whole-genome duplication (WGD) events, as well as the evolution of secondary metabolite clusters and resistance genes. These results improved the understanding of the genomic architecture of *A. indica*. This chromosome-level genome assembly can be used as a new reference genome for *A. indica*, laying a substantial foundation for further genomic studies.

MATERIALS AND METHODS

Plant Material, DNA Preparation, and Genome Sequencing

Fresh tissues of *A. indica* were randomly collected from a locally grown tree in the Liufang Yuan park of Hainan University (100.61438 E, 36.28672 N), Hainan Province, China. Fresh leaves were collected to isolate genomic DNA of *A. indica* for *de novo* sequencing and assembly. Genomic DNA was extracted from leaves of *A. indica* using the DNAsecure Plant Kit (TIANGEN, Biotech Co., Ltd., Beijing, China). For Illumina sequencing, a paired-end library with an insert size of 270 bp was generated and sequenced on the Illumina HiSeq X Ten platform. For PacBio sequencing, a 20 kb insert library was generated and sequenced on the PacBio RSII platform.

Genome Assembly

First, Canu v2.0 (Koren et al., 2017) software was used to correct and assemble raw PacBio sequencing reads, and 886 contigs with N50 ∼ 6 M were assembled by Canu. In addition, we performed a round of polishing on the assembled contigs using the RACON (Vaser et al., 2017) with the PacBio long reads, and the polished contigs were further corrected two rounds on the genome-wide base-level by Pilon v1.21 (Walker et al., 2014) with the Illumina short reads. 870 contigs were left after error correction with RACON and Pilon software. Genome size of *A. indica* was estimated by flow cytometry (Pellicer and Leitch, 2020).

Chromosome Assembly Using Hi-C

For Hi-C sequencing, a 150 bp paired-end library was generated and sequenced on the Illumina HiSeq X Ten platform. Bowtie2 (Langmead and Salzberg, 2012) with the default parameters was used to map the clean reads to the *A. indica*. HiC-Pro v2.11.1 (Servant et al., 2015) was used to map the Hi-C sequencing reads to the assembled draft genome and detect the valid contacts. Then we used ALLHIC v0.9.12 (Zhang et al., 2019) to cluster contigs into chromosome-scale scaffolds based on the relationships among valid contacts.

Assessment of Genomic Integrity

The draft genome sequence of *A. indica* (GCA_000439995.3) was downloaded from NCBI as a reference. The accuracy and integrity of the genome assembly was evaluated using BUSCO v3.0.2, based on the OrthoDB[1] database. LTR Assembly Index (LAI) scores were calculated by LTR_Retriever (v2.8) with the default parameters (Ou and Jiang, 2018; Ou et al., 2018). The transcriptomic NGS short reads from 5 tissues of *A.indica* (SRR12709585, SRR12709584, SRR12709583, SRR12709582, and SRR12709581) (Wang et al., 2020) were mapped against the assemblies using Hisat2 (Kim et al., 2015) with default parameters. The genomic NGS short reads were also mapped to the assemblies using Bowtie2. Finally, the collinearity analysis between our assembly and GCA_000439995.3 was performed with Minimap2[2] and dotPlotly.[3]

Repetitive Elements

We identified repetitive elements through both RepeatModeler v1.0.8 (Price et al., 2005) and RepeatMasker v4.0.7 (Tarailo-Graovac and Chen, 2009). The LTRs of *A. indica* were identified by using LTRharvest v1.6.1 (Ellinghaus et al., 2008) and

[1]http://cegg.unige.ch/orthodb

[2]https://github.com/galaxyproject/tools-iuc/tree/master/tools/minimap2

[3]https://github.com/tpoorten/dotPlotly

LTR_Finder v1.05 (Xu and Wang, 2007). LTR_retriever v2.8.7 (Ou and Jiang, 2018) was used to integrate the results of LTRharvest and LTR_Finder. RepeatModeler employed RECON v1.08 and RepeatScout v1.0.5 to predict interspersed repeats and then combined the repeat sequences from LTR-retriever with the repeat sequences from RepeatModeler to be the local repeat library. To recover the repeats in the *A. indica* genome, a homology-based repeat search was conducted by using RepeatMasker with the *ab initio* repeat database and Repbase.[4]

Non-coding RNAs

Non-coding RNAs were detected through searching against various RNA libraries. Reliable tRNA positions were searched *via* tRNAscan-SE v1.3.1 (Lowe and Eddy, 1997). Small nuclear RNAs (snRNAs) and microRNAs (miRNAs) were searched by using INFERNAL v1.1 (Nawrocki and Eddy, 2013) against the Rfam (Griffiths-Jones et al., 2005) database.

Gene Prediction

Homology annotation was performed using genomes of three representative species, including *Citrus sinensis* (Xu et al., 2013), *Theobroma cacao* (Argout et al., 2011), and *Acer yangbiense* (Yang et al., 2019). The TBLASTN software (Camacho et al., 2009) was used to align the protein sequences of these species to *A. indica* genome sequence, with an *E*-value ≤ 1e-5. The exact gene structures were predicted using GeneWise 2.2.0 (Birney et al., 2004) according to the TBLASTN results. We used Cufflinks v2.2.1 (Trapnell et al., 2012) to preliminarily identify gene structures based on the RNA-seq data. *ab initio* annotation was performed using Augustus v3.2.2 (Stanke et al., 2004) and SNAP (Korf, 2004) with the repeat-masked genome sequences. All genes predicted from the three annotation procedures were integrated with MAKER (Holt and Yandell, 2011) software.

Functional Annotation

The protein sequences of the consensus gene set were aligned to four protein databases, including NR,[5] InterPro,[6] Swiss-Prot,[7] and EggNOG (Powell et al., 2012), for predicted gene annotation. The physically clustered specialized metabolic pathway genes were identified by the PlantiSMASH analytical pipeline (Kautsar et al., 2017). Plant disease resistance (R) genes were predicted by the Disease Resistance Analysis and Gene Ontology (DRAGO) pipeline (Osuna-Cruz et al., 2018).

Phylogenetic Analysis and Expansion/Contraction of Gene Families

The genome of *A. indica* and 13 other plants were selected for phylogenetic analysis. All-vs.-all BLASTP (Altschul et al., 1997) search results with an *E*-value ≤ 1e-5 were grouped into orthologous and paralogous clusters using OrthoFinder v2.3.7 (Emms and Kelly, 2019). Multiple sequence alignments of all single-copy orthologous gene families were performed by using MUSCLE (Edgar, 2004). The set of single nucleotide polymorphisms (SNPs) presented in each single-copy orthologous gene family was extracted and then integrated according to the arrangement of the genes on the *A. indica* genome. A maximum likelihood (ML) tree was constructed using the integrated SNPs by PhyML v3.1 (Guindon et al., 2009). Divergence time between species was estimated using MCMCtree, which was incorporated in the PAML v4.8 package (Yang, 1997). CAFÉ v3.1 (De Bie et al., 2006) was used to measure the expansion/contraction of orthologous gene families.

Genome Duplication Analysis

MCScan v0.8 (Tang et al., 2008) package with default parameters was used for the detection of syntenic blocks, defined as regions with more than 5 collinear genes. We aligned the amino acid sequences of syntenic block gene pairs and reciprocal best hits (RBH) gene pairs using MAFFT and further aligned their nucleotide sequences using ParaAT (Zhang et al., 2012). The synonymous substitution rate (*Ks*) values of these gene pairs were calculated using YN model in KaKs_Calculator v2.0 (Wang et al., 2010). The value of *Ks* peak was determinated by the abscissa value of the highest point of the *A. indica Ks* plot. The WGD events of each species were estimated based on the *Ks* distributions. The gene pairs with the median $Ks < 0.05$ were defined as the retained genes from the recent segmental duplication. According to the formula $T = Ks/2r$, the *Ks* values were converted to divergence times, where T is divergence time and r is the neutral substitution rate ($r = 3.39 \times 10^{-9}$). The paralog analysis in *A. indica* genome were performed using RBH from all-vs.-all BLASTp searches using *A. indica* protein sequences. RBHs are defined as reciprocal best BLASTp matches with *e*-value threshold of 1e-5, c-score threshold of 0.3 (Guo et al., 2018).

Identification and Phylogenetic Analysis of Terpene Synthase and Cytochrome P450 Family Members

Genomes were aligned using HMMER 3.0 search with an *E*-value 1e-5 against the Pfam-A database (02-May-2020) locally. PF01397 (Terpene synthase, N-terminal domain) and PF03936 (Terpene synthase family, metal binding domain) domains were used to identify the members of the TPS gene family. The collection used for phylogenetic analysis consisted of 403 putative TPSs from *A. indica* and other 13 plants and six reported TPSs belonged to TPS- a (AAX16121.1), b (AAQ16588.1), c (AAD04292.1), e (Q39548.1), f (Q93YV0.1), and g (ADD81294.1) subfamilies (Kumar et al., 2018b; Zhou et al., 2020). PF00067 (Cytochrome P450) was used to identify the members of the CYP gene family. Putative CYPs were screened by amino acid length ($450 < \text{length} < 600$) to perform phylogenetic analysis. Protein sequences were aligned using ClustalX in MEGAX using default sets (Kumar et al., 2018a). The ML trees were constructed based on the alignment of TPS and CYP protein sequences using MEGAX software with 100 bootstrap replicates, respectively. The identification of *A. indica*-specific TPS and CYP genes was based on the phylogenetic analysis using other 13 plant genome as the

[4]https://www.girinst.org/repbase/

[5]https://www.ncbi.nlm.nih.gov/protein/

[6]https://www.ebi.ac.uk/interpro/

[7]http://www.uniprot.org

outgroup and a cutoff of 55% identity, which indicated separate subfamily assignment (Liu et al., 2018; Tu et al., 2020).

RESULTS

Genome Sequencing and Assembly

To obtain a chromosome-level assembly of *A. indica*, the genome was sequenced using a combination of Illumina, PacBio, and Hi-C methods, and assembled by a hierarchical approach. A total of 110 Gb (providing 188 × genome coverage) Illumina paired-end short reads were produced and the heterozygosity ratio was estimated to be 0.896%. Based on the 21-mer depth distribution of the Illumina short reads, the genome size was estimated to be 165 Mb (**Supplementary Figure 1**).

We also generated 126 Gb of raw PacBio sequencing reads from the single-molecule real-time (SMRT) sequencing platform, reaching 256 × coverage of the *A. indica* genome (**Supplementary Figure 2** and **Supplementary Table 1**). The total size of the reads assembled from the post-correction genome was 281,629,231 bp with a GC content of 32.2%, consisting of 870 contigs. The contig N50 was 6,039,544 bp, and the longest contig was 15,111,501 bp. Our genome assembly constitutes ∼73.2% of the 385 Mb genome estimated by flow cytometry (Pellicer and Leitch, 2020).

We further conducted the Hi-C sequencing to scaffold the preliminary assemblies and enhance the assembled contiguity at the chromosome level. In total, the Hi-C sequencing generated approximately 40.48 Gb clean reads. 94.5% reads from Hi-C sequencing were mapped to the assembled contigs, of which 26.1% were unique mapped read pairs (**Supplementary Table 1**). The verified read pairs were selected after considering the map position and orientation of the unique mapped read pairs. Then, according to the contiguity information between Hi-C read pairs, ALLHIC software was used to cluster, order, and orient the previous assemblies for chromosome-level scaffolding (**Figure 1A**). A total of 70 scaffolds were obtained after Hi-C sequencing reads assist chromosome assemble, of which 14 scaffolds formed chromosomes (**Figure 1B** and **Supplementary Table 1**). The final size of the *A. indica* genome assembly was 281 Mb, and the scaffold N50 was 19 Mb (**Table 1**).

Evaluation of the Genome Assembly

The quality of the assembly was assessed and compared with the reference genome sequence of *A. indica* from NCBI (GCA_000439995.3) (**Supplementary Figure 3** and **Table 2**). The Benchmarking Universal Single-Copy Orthologs (BUSCO) (Simao et al., 2015) analysis was used to evaluate the integrity of the genome. The BUSCO assessment showed that the completeness of the assembled genome of *A. indica* was 91.7%, which was much higher than that of the reference genome (**Supplementary Figure 3A**, **Table 2**, and **Supplementary Table 2**). The average LAI score of *A. indica* genome was 4.82, which was lower than the "reference" quality ($10 < LAI < 20$) based on the LAI classification (Ou et al., 2018). The Illumina short reads were also used to assess the integrity of the genome. The transcriptomic Illumina sequencing short reads were mapped to the two assemblies by Hisat2 (Kim et al., 2015), and approximately 92.76 and 87.49% of the reads were mapped to our assembly and GCA_000439995.3, respectively. By using Bowtie2 (Langmead and Salzberg, 2012) software, the genomic Illumina sequencing short reads were also mapped to the assemblies. About 99.29 and 97.09% of the Illumina short reads could map to our assembly and GCA_000439995.3, respectively (**Supplementary Figure 3B**). Finally, collinearity analysis revealed good collinearity between our assembly and GCA_000439995.3 (**Supplementary Figure 3C**).

Gene Prediction and Genome Annotation

Gene models were generated by a combination of reference plant protein homology support, transcriptome data, and *ab initio* gene prediction. All gene models were merged with MAKER (Holt and Yandell, 2011), resulting in a total of 25,767 protein-coding genes with an average sequence length of 2,837 bp. On average, each predicted gene contained 5.4 exons with a mean sequence length of 231 bp (**Table 1**). In addition, 3,856 non-coding RNAs, including 1,381 rRNAs, 1,204 tRNAs, 173 microRNAs (miRNAs), and 1,098 small nuclear RNAs (snRNAs) were identified (**Supplementary Table 3**). We also identified 40.99% of the assembled sequences as repetitive sequences, which was higher than that of the reported genomes (Kuravadi et al., 2015; Krishnan et al., 2016). The majority of the repeats were long terminal repeats (LTRs), constituting 16.88% of the genome. Unclassified elements, DNA elements, and long interspersed nuclear elements (LINEs), accounted for 14.28, 6.54, and 1.08% of the genome, respectively (**Supplementary Table 4**).

To further evaluate the functional validity of the predicted genes, Diamond, BLASTP, InterProScan and EggNOG-mapper were utilized by searching the Nr, SwissProt, InterPro, and EggNOG databases (**Supplementary Figure 3D**). Overall, 24,801 genes (96.2%) were functionally assigned. 95.4 and 81.6% of these genes found homologies and annotated proteins in the Nr and SwissProt databases, respectively. 84.3% of the genes were detected with conserved protein domains using InterProScan. In addition, 47.4% of the genes were categorized by Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway (Moriya et al., 2007; **Supplementary Table 5**).

Phylogenetic Analysis

To investigate the genetic diversity and evolutionary history of *A. indica* genome, a gene family clustering analysis with the *A. indica* genome and 13 other representative plant species was performed. These selected species included two plants in the *Sapindales* order (*Acer yangbiense* and *Citrus sinensis*), eight plants in the eudicot clade (*Arabidopsis thaliana*, *Theobroma cacao*, *Gossypium raimondii*, *Carica papaya*, *Vitis vinifera*, *Cucumis sativus*, *Fragaria vesca*, *Prunus persica*, and *Solanum lycopersicum*), and two outgroup species (*Brachypodium distachyon* and *Amborella trichopoda*).

OrthoFinder (Emms and Kelly, 2019) was used to construct a phylogenetic tree with 1,338 single-copy orthologous genes among 14 species, which showed that *A. indica* is most closely related to *C. sinensis* (**Figure 2**). Further analysis showed that 36 gene families were specific to *A. indica* (**Supplementary Table 6**).

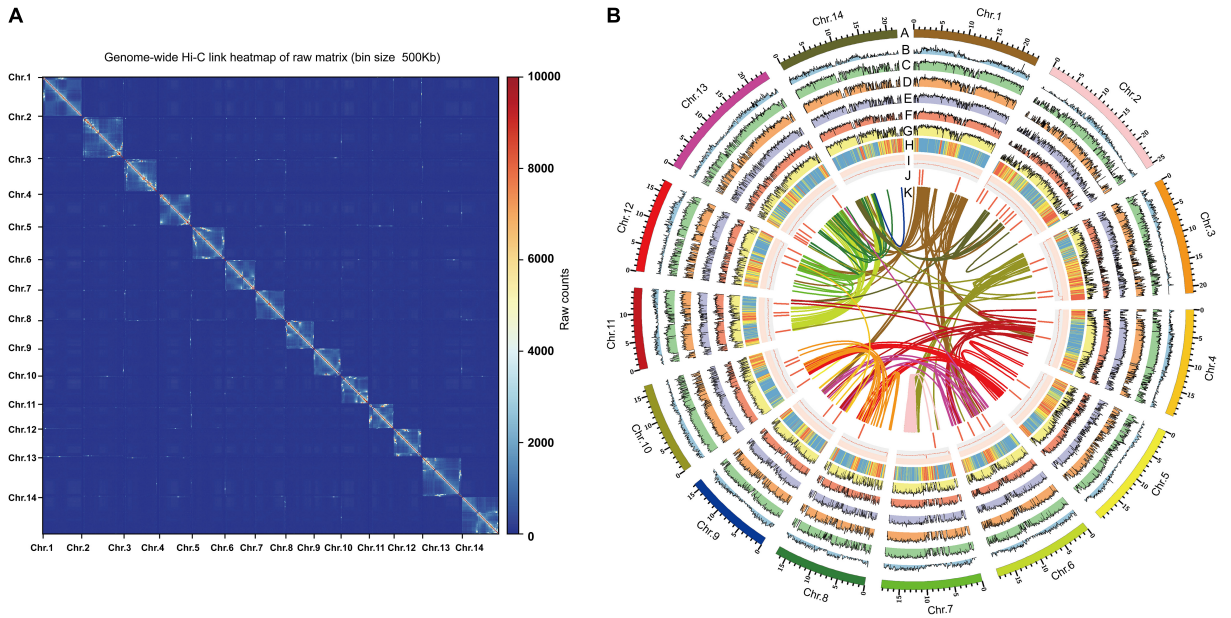

FIGURE 1 | Genome features of genome assembly of *A. indica*. **(A)** Hi-C contact data mapped on the *A. indica* genome showing genome-wide all- by-all interactions. **(B)** The landscape of genome assembly and annotation of *A. indica*. (A) Circular representation of the pseudomolecules. (B) The distribution of gene density with densities calculated in 500 kb windows. (C–G) Expression of *A. indica* genes (from outside to inside tracks: stem, root, leaf, fruit and flower). (H,I) The distribution of repeat density and GC density with densities calculated in 500 kb windows. (J) Locations of genes mapped to secondary metabolism. (K) Syntenic blocks.

Enrichment analysis showed that these specific genes were mostly involved in "binding," "catalytic activity," "metabolic process," "cellular process," and "membrane" (**Supplementary Table 7**). With the divergence time between *P. persica* and *F. vesca* as a calibration point (with the corrected time obtained from TimeTree Kumar et al. (2017)), the divergence time among these species were also estimated. *A. indica* and *C. sinensis* diverged from a common ancestor ∼57 Mya (**Figure 2**). To better understand the genetic basis of *A. indica*, the expansion and contraction of gene families were investigated. 997 gene families were expanded in *A. indica*, while 293 gene families were contracted from the *A. indica* genome. Compared with *C. sinensis*, which has 369 expanded gene families and 682 contracted gene families, *A. indica* has expanded more gene families. GO and KEGG analysis of the expanded and contracted gene families were also performed (**Supplementary Figure 4** and **Supplementary Table 8**). The *A. indica*-specific expanded and contracted gene families might be related to the adaptation to *A. indica*-specific tropical niches. Further researches are required to verify the function of these genes.

Genome Duplication Analysis

To investigate genome wide duplications in *A. indica* genome, self-comparison of the *A. indica* genome was performed using MCScan (Tang et al., 2008; **Supplementary Figure 5**). 242 homologous blocks were identified in the intragenomic gene synteny of *A. indica*, containing 2,281 gene pairs. These homologous blocks were distributed across the 14 chromosomes, covering 17.66% of protein-coding genes (4,139/25,767). The

TABLE 1 | Statistics of the *A. indica* genome assembly.

Feature	Value
Genome size (Mb)	281
Genome GC%	32.2
N50 (Mb)	19
Gene number	25,767
Average gene length (bp)	2,837
Exon no. per gene	5.4
Exon number	138,941
Average exon length (bp)	231
Total exon length (bp)	32,191,037

synonymous nucleotide substitutions (*Ks*) of the gene pairs peaked at approximately 0.01 and 1.12 (**Figure 3A**). The first peak at approximately 1.12 indicated the core eudicot γ triplication event (∼165 Mya). The second peak at approximately 0.01 indicated a relatively recent duplication event or events. To distinguish whether this peak represents a whole genome duplication event or background duplications, we performed synteny analysis on *A. indica*, *V. vinifera*, *C. sinensis,* and *A. yangbiense* genomes (**Figure 3B** and **Supplementary Figure 6**). Intergenomic collinearity analysis showed 611 homologous blocks containing 14,674 gene pairs and a 3:3 syntenic relationship between *A. indica* and *V. vinifera* (**Figure 3B** and **Supplementary Figure 7A**). Although there were 2:1 syntenic relationship between *A. indica* vs. *C. sinensis* and *A. indica* vs. *A. yangbiense* (**Supplementary Figures 7B,C**),

TABLE 2 | Comparison of the *A. indica* genome assembly versions.

Feature	This study	Krishnan et al., 2012	GCA_000439995.3
Sequence technology	Illumina + PacBio + Hi-C	Illumina + PacBio	Illumina
Assembly level	Chromosome	Scaffold	Contig
Genome size (Mb)	281	216	264
Genome GC%	32.2	31.9	32.0
Number of scaffolds	70	25,560	126,142
Scaffold N50 (bp)	19,542,739	2,629,187	3,491
Number of contigs	870	48,555	142,701
Contig N50 (bp)	6,039,544	25,406	3,310
BUSCO	91.7%	91.4%	79.9%

only 13 and 14% of the *A. indica* gene models in syntenic blocks, respectively, were present as two copies. Meanwhile, we did not identify large *C. sinensis* and *A. yangbiense* segments that have two syntenic copies in *A. indica* by the synteny dot plot of *A. indica* vs. *C. sinensis* and *A. indica* vs. *A. yangbiense* (**Supplementary Figure 6**). Our analysis indicated that *A. indica* didn't experience additional WGD after the γ event, but a recent small-scale segmental duplication (Xu et al., 2013; Yang et al., 2019). The calculation of *Ks* for *A. indica* vs. *C. sinensis* indicated that this recent segmental duplication event occurred approximately 1.5 Mya. Furthermore, we also performed paralog analysis in *A. indica* genome using reciprocal best hits (RBH) from primary protein sequences by all-vs.-all BLASTp matches. We detected 6,298 RBH paralogous gene pairs in the *A. indica* genome, and the RBH paralog *Ks* distribution shows a *Ks* peak at around 0.01 (**Supplementary Figure 8**). That this RBH *Ks* peak is close to the syntelog *Ks* peak also indicates *A. indica* has a recent segmental duplication mixed with gene duplication.

Generally, gene duplication events vary the genomic architecture, including genome size, genome density, gene content, and gene expression. In this study, we defined the RBH paralogous gene pairs with the median $Ks < 0.05$ as the retained genes from recent gene duplication. A total of 768 gene pairs were retained after recent gene duplication. GO analysis revealed that these gene pairs were significantly involved in binding, catalytic activity, metabolic process, cellular process, and reproductive process (**Figure 3C**). Recent gene duplication may also affect the percentage of genes in many function categories with different contributions. In the *A. indica* genome, the percentage of retained genes from recent gene duplication in "catalytic activity GO:0003824," "recognition of pollen GO:0048544," "pollen-pistil interaction GO:0009875," "pollination GO:0009856," and "multi-multicellular organism process GO:0044706" was greater than that of the average genome content (**Figure 3C** and **Supplementary Figure 9**). We further calculated the omega values (*Ka*/*Ks*) for most of the homologous gene pairs. Most of the omega values for the homologous gene pairs were smaller than 1, which indicated that purifying selection may be the predominant action within the retained genes from recent gene duplication (Stix, 1992). However, 120 gene pairs were identified that have experienced potential positive selection. GO analysis showed that these genes were mainly enriched in "catalytic activity, acting on a protein GO:0140096," "protein binding GO:0005515," "protein-containing complex GO:0032991," and "membrane-bounded organelle GO:0043227" (**Supplementary Table 9** and **Supplementary Figure 10**).

Secondary Metabolite Analysis

Genes encoding some specialized metabolic pathways are found physically clustered in plant genomes (Nutzmann et al., 2016; Liu et al., 2020). We utilized the PlantiSMASH analytical pipeline (Hu et al., 2019) to identify physically clustered specialized metabolic pathway genes. According to the analysis, 50 clusters including 692 genes were identified in the *A. indica* genome (**Supplementary Table 10**). The sizes of the identified clusters range from 27.2 to 1634.4 kb. 105 (out of 692) clustered genes were contained in the 997 *A. indica*-specific expansion gene families (**Supplementary Table 10**). Furthermore, 41 (*C. sinensis*), 51 (*A. yangbiense*), 48 (*T. cocoa*), 47 (*G. raimondii*), 45 (*A. thaliana*), 35 (*F. vesca*), 33 (*P. persica*), 30 (*C. sativus*), 46 (*V. vinifera*), 47 (*S. lycopersicum*), and 29 (*B. distachyon*) clusters were detected in other 11 species (**Figure 4A**). As expected, more terpene-related clusters were identified in the *A. indica* genome than that of other species.

Azadirachtin is a triterpenoid compound of neem tree, which has effective insecticidal activities against a wide range of insect species, but has very low toxicity to mammals. Terpene synthase (TPS), cytochrome P450 (CYP450), alcohol dehydrogenase (ADH), acyltransferase (ACT), and esterase (EST) were proposed to be involved in biosynthesis of azadirachtin (Wang et al., 2020). In this study, a large number of genes encoding CYP 450s (78), TPSs (58), and ACTs (34) were identified in secondary metabolite biosynthesis gene clusters. Genes encoding ADHs, and ESTs may reside dispersed in the genome. The terpene-related clusters mainly distributed on chromosome 1, 2, 3, 5, 6, 7, 10, 11, 12, and 13. Four terpene-related clusters (cluster 18–21) covering ∼1.4 Mb were distributed on chromosome 13 (**Figure 4B**). Among the 83 clustered terpene-related genes on chromosome 13, 12 genes were contained in the *A. indica*-specific expanded gene families. These genes are proposed to be potential genes participated in the terpene biosynthesis specific to *A. indica*.

KEGG enrichment analysis was performed to investigate the function of genes on chromosome 13. The result showed that genes on chromosome 13 were mainly involved in "Protein processing in endoplasmic reticulum," "Sesquiterpenoid and triterpenoid biosynthesis," and "Ovarian steroidogenesis" (**Figure 4C**). Furthermore, when we performed GO enrichment analysis using the genes on chromosome 13, the genes associated with "terpene synthase activity" (GO:0010333) exhibited a low *P*-value, indicating that "terpene synthase activity" was the most enriched functional category of chromosome 13 (**Figure 4D**).

Terpene Synthase Gene Family

TPS gene family is characterized by two large domains: PF01397 (Terpene synthase, N-terminal domain) and PF03936 (Terpene synthase family, metal binding domain). To investigate the characteristics and evolution of the TPS gene families, we

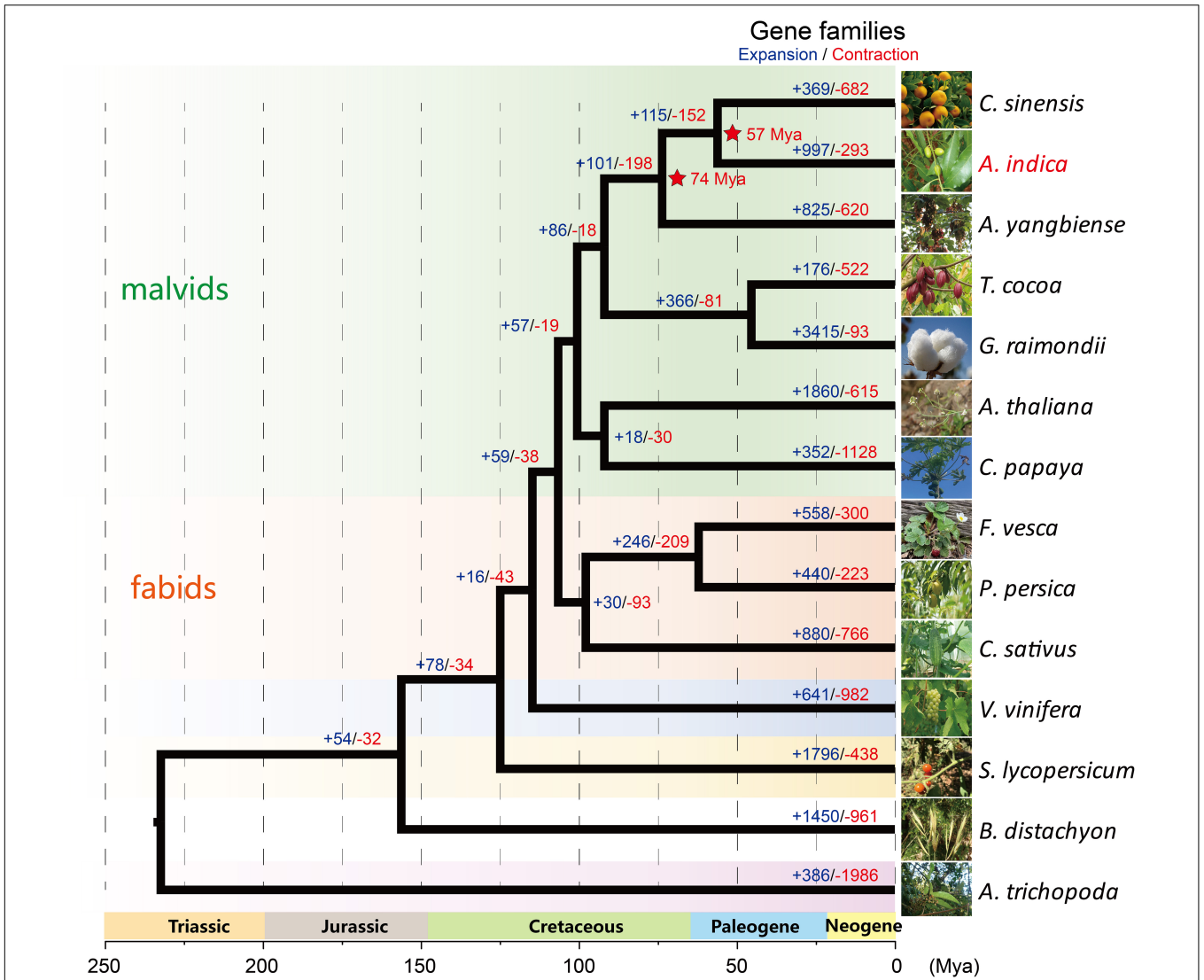

FIGURE 2 | Maximum Likelihood based phylogenetic analysis of *A. indica* and 13 other plant species using 1,338 single-copy orthologous genes with PhyML software (Guindon et al., 2009). The estimation of the divergence times (Mya; red; star) of *A. indica* with its closely related species are indicated. The number of expanded gene families (+; blue) and the number of contracted gene families (–; red) are shown in each branch.

identified a total of 512 putative TPS genes in *A. indica* and other 13 plant genome. 70 putative TPS genes were identified in *A. indica*; These consisted of 44 AziTPS genes containing both PF01397 and PF03936 domains, nine AziTPS genes containing PF01397 domain, and 17 AziTPS genes containing PF03936 domain. *A. indica* (N = 70) contained the most copies of TPSs compared with other plants, followed by *C. sinensis* (N = 49) and *A. yangbiense* (N = 57) (**Figure 5A** and **Supplementary Table 11**). In addition, eight AziTPS genes experienced recent gene duplication (**Supplementary Table 11**).

Phylogenetic analysis was performed using 403 TPSs (the remaining TPS genes were too short for meaningful alignment) from *A. indica* and other 13 plants, including six reported TPS genes belonged to TPS- a, b, c, e, f, and g subfamilies, respectively (**Supplementary Table 11**). As shown in **Figure 5B**, the topology of six subfamilies is similar to that of the previous papers (Kumar et al., 2018b; Zhou et al., 2020; Ji et al., 2021). Among the 40 AziTPS used in phylogenetic analysis, 15, 13, 4, 3, 1, and 4 AziTPS genes fell in TPS- a, b, c, e, f, and g subfamilies, respectively. TPS-a and -b subfamilies were the main subfamilies in *A. indica*, approximately 37.5 and 32.5% of the total AziTPS genes in phylogenetic analysis. This is in accordance with other plant species, including tea, grape, and Chinese mahogany (Chen et al., 2011; Zhou et al., 2020; Ji et al., 2021). Furthermore, we identified putative *A. indica*-specific TPSs using phylogenetic analysis and a cutoff of 55% identity, which indicates separate subfamily assignment (Liu et al., 2018; Tu et al., 2020). A total of nine *A. indica*-specific TPS genes were identified (**Supplementary Table 11**). Interestingly, seven of these specific AziTPSs (Indica_007028, Indica_007047,

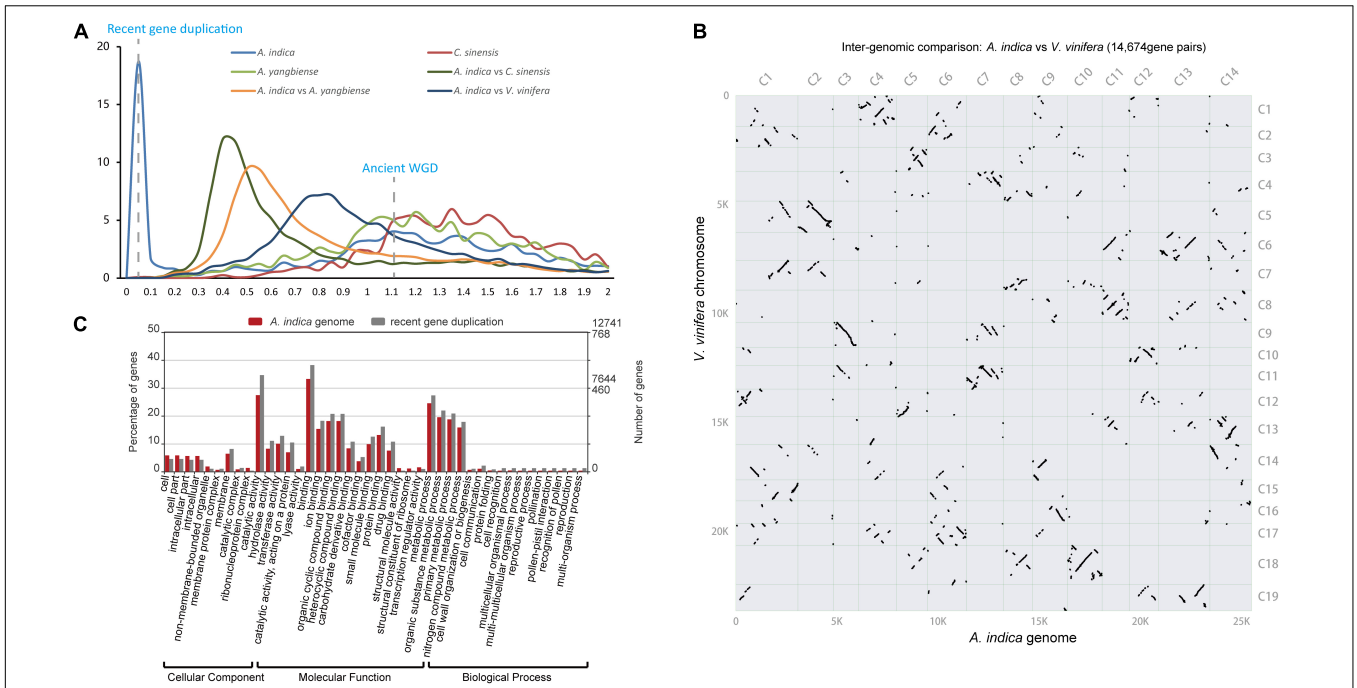

FIGURE 3 | Genome evolution of *A. indica*. **(A)** *Ks* distribution (The gene pairs located in syntenic blocks) between syntenic genes within the *A. indica* genome or between genomes. **(B)** Inter-genomic syntenic analysis between *A. indica* genome and *V. vinifera* genome. **(C)** GO enrichment of the *A. indica* genome (green) and genes retained after recent gene duplication (orange) ($P < 0.05$). *Pearson chi-square test *P*-value < 0.05; **Pearson chi-square test *P*-value < 0.01.

Indica_007053, Indica_007068, Indica_007070, Indica_007072, and Indica_007143) were located in the terpene-related clusters (cluster 18, 19, and 20) of chromosome 13 (**Supplementary Table 11**).

We further investigated the expression pattern of TPS genes in *A. indica*. Transcriptome datasets from five tissues of *A. indica* were obtained from our previous study (Wang et al., 2020) and remapped to the chromosome-level genome assembly in this study. More than 88% of the RNAseq reads were mapped uniquely to the genome assembly across all samples (**Supplementary Table 12**). Transcripts of 27 TPS genes were detected in the tested tissues. Most of the detected transcripts exhibited a spatial-specific expression pattern (**Figure 5C**). Nine, one, four, three, and four genes were exclusively expressed in flower, fruit, root, leaf, and stem, respectively. Seven genes (AziTPS30, −48, −5, −57, −26, −63, and −50) were primarily expressed in one or two tissues.

Cytochrome P450 Gene Family

The characteristics and evolution of the cytochrome P450 (CYP) gene families were also investigated. In total, 3,657 CYP genes were identified from all 14 plant genomes (**Figure 6A** and **Supplementary Table 13**). A total of 355 CYP genes were in the *A. indica* genome, of which 36 CYP genes were involved in recent gene duplication (**Supplementary Table 13**). Moreover, 157 full length CYP (450 < length < 600) protein sequences of *A. indica* were aligned to construct a phylogenetic tree. As shown in **Figure 6B**, the phylogenetic tree was divided into two major clades: A type (49%; 77/157) and non-A type (51%; 80/157); and further clustered into nine clans. The Clan 71 is the largest clan and comprises of 49% (77/157) members; 18, 4, 28, and 25 members are classified into Clan72, Clan74, Clan85, and Clan86; remaining Clan51, Clan710, Clan711, and Clan727 are single family clans.

In order to identify putative *A. indica*-specific CYP genes, we constructed a phylogenetic tree using amino acid sequence alignment of 2,807 (450 < length < 600) CYP genes in *A. indica* and other 13 plants genome with a cutoff of 55% identity (Liu et al., 2018; Tu et al., 2020). Six *A. indica*-specific CYP genes were identified (**Supplementary Table 13**). Similar to TPS genes, five of these CYP genes (Indica_007272, Indica_007273, Indica_007276, Indica_007277, and Indica_007278) were located in the terpene-related cluster 21 of chromosome 13 (**Supplementary Table 13**). These specific-TPSs and CYPs in the terpene-related secondary metabolite biosynthesis gene clusters of chromosome 13 might be involved in the specific biosynthesis of azadirachtin.

We also investigated the expression pattern of *A. indica* CYP genes in different tissues (fruit, flower, root, stem, and leaf). Transcripts of 221 CYP genes were detected with different patterns (**Figure 6C**). There were more high-expressed CYPs in fruit, stem and leaf than flower and root. The high-expressed CYPs in fruit, stem and leaf were 83, 88, and 97, respectively. CYPs with a high-expression in the tissues (fruit and leave) with high azadirachtin. A content, are more likely to be involved in azadirachtin biosynthesis. Furthermore, *A. indica*-specific AziCYP256 (Indica_007272) and AziCYP8 (Indica_007273) were highly expressed in fruit and flower.

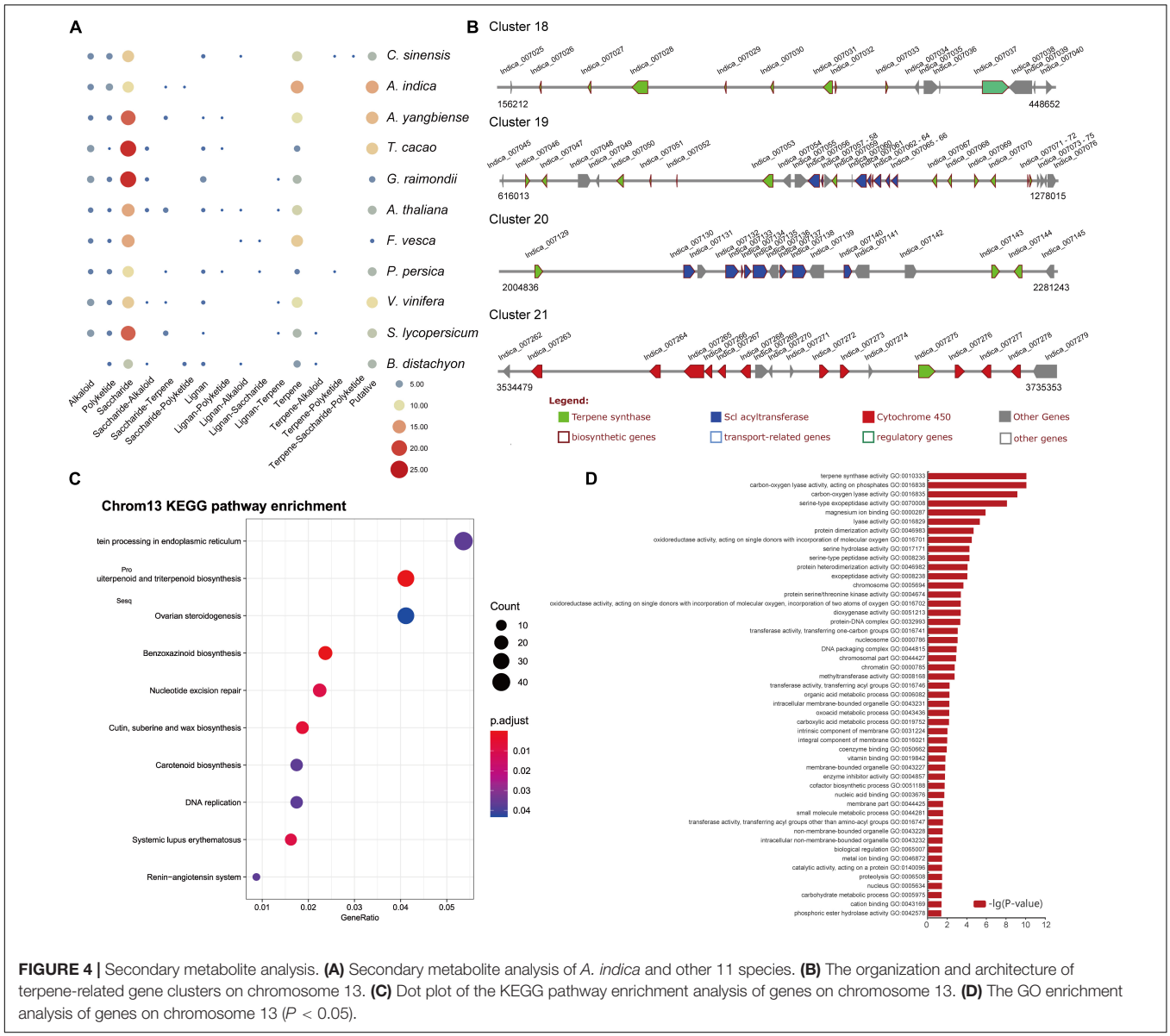

FIGURE 4 | Secondary metabolite analysis. **(A)** Secondary metabolite analysis of *A. indica* and other 11 species. **(B)** The organization and architecture of terpene-related gene clusters on chromosome 13. **(C)** Dot plot of the KEGG pathway enrichment analysis of genes on chromosome 13. **(D)** The GO enrichment analysis of genes on chromosome 13 ($P < 0.05$).

Resistance Genes

Plants have developed a wide range of defense mechanisms to protect themselves against the attack of pathogens in their constant struggle for survival. In general, proteins encoded by resistance (R) genes display modular domain structures. In this study, putative R genes in the *A. indica* genome (1,488) and other 13 species were identified (**Supplementary Table 14**). In the *A. indica* genome, 238 R genes may exert their disease resistance function as cytoplasmic protein through canonical resistance domains, such as the nucleotide-binding sites (NBSs), the leucine-rich repeat (LRR), and terminal inverted repeat (TIR) domains (**Supplementary Table 14**). 167 NBS genes were identified in the *A. indica* genome, which could be divided into five classes according to the conserved domains: N, CN, CNL, NL, and TNL. The majority were N type which contained only the NB-ARC domain. In comparison with other genomes in malvids, most of the NBS genes in the *A. indica* genome were underrepresented relative to other Sapindales genomes (*C. sinensis* and *A. yangbiense*) and Malvales genomes (*T. cacao* and *G. raimondii*), but overrepresented relative to other Brassicales genome (*A. thaliana* and *C. papaya*). In addition, 447 genes were classified as transmembrane receptors, including 221 receptor-like kinases (RLK), and 226 receptor-like proteins (RLP). 721 putative kinases were also identified in the *A. indica* genome.

DISCUSSION

A. indica is a valuable plant species given its economic and pharmaceutical significance (Stix, 1992). A high-quality reference genome is essential for the genetic and genomic studies of

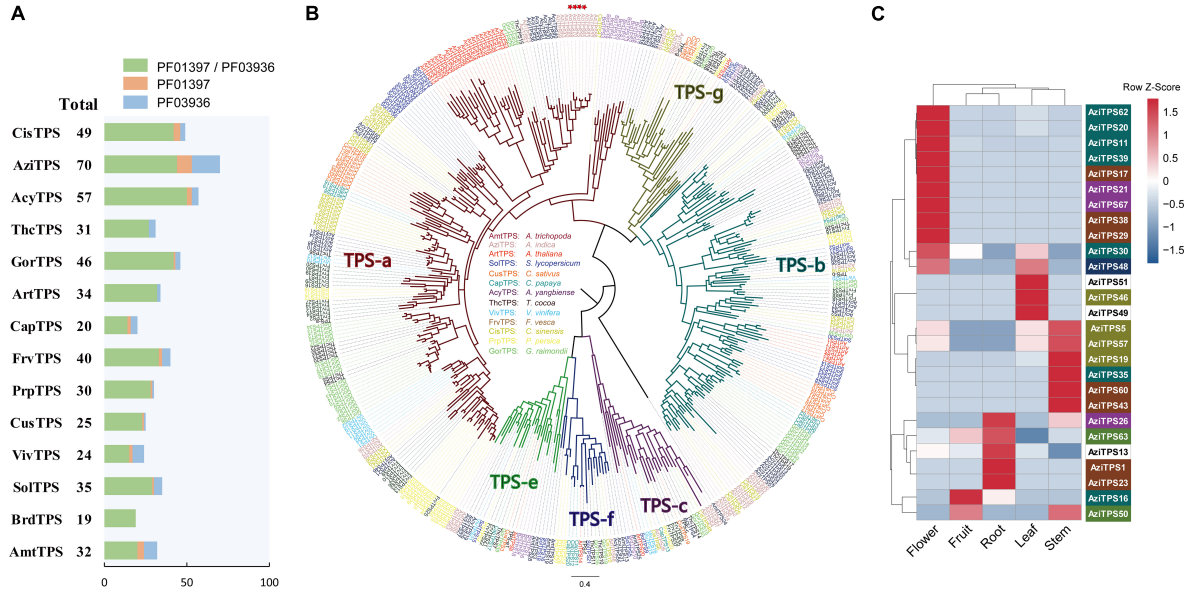

FIGURE 5 | Analysis of TPS gene family in *A. indica*. **(A)** Statistics of predicted TPS genes in *A. indica* and other 13 plants. **(B)** Phylogenetic analysis of TPS family members based on the protein sequences. A Maximum-Likelihood (ML) tree was generated from an alignment of 409 TPS protein sequences, comprising 403 putative TPSs from *A. indica* and other 13 plants (the remaining TPS genes were too short for meaningful alignment) and other six reported TPS genes belonged to TPS- a (AAX16121.1), b (AAQ16588.1), c (AAD04292.1), e (Q39548.1), f (Q93YV0.1), and g (ADD81294.1) subfamilies. The *A. indica* TPS genes are highlighted by red stars. **(C)** *A. indica* TPS genes substantially expressed in five tested organs (root, flower, fruit, leaf, and stem). The color of TPS genes depends on the subfamily information.

A. indica. However, molecular-level studies on this species are limited. Here, we assembled the first chromosome-scale genome of *A. indica* by a combination of Illumina, PacBio, and Hi-C technology. The size of the genome assembly is approximately 281 Mb, with a scaffold N50 value of 19 Mb. The N50 of our assembled genome is much higher than that of the previous published draft genomes (Krishnan et al., 2012, 2016; Kuravadi et al., 2015). Our assembled genome size covered ∼73.2% of the estimated genome size (385 Mb) by flow cytometry. However, previously assembled 12 contig-level *A. indica* genomes were generally less than 300 Mb (Krishnan et al., 2016). The *A. indica* genome shows a high level of heterozygosity (0.896%) and repeat content (40.99%), rendering substantial challenges for its assembly (Nowak et al., 2015). Hi-C technology has been broadly available for many complex species (Chen et al., 2020). In this study, Hi-C technology facilitated the completeness and accuracy of a chromosome-level genome assembly for *A. indica*. The improvement of BUSCO evaluation shows that our assembly represents a better template for gene annotation than the reference sequence. Considering that the genome is highly heterozygous and repetitive, the present version represents a high-quality genome assembly. The obtained genome is also the second chromosome-level genome of the Meliaceae family, which will pave the way for further genetic and genomic studies of this family.

Gene duplication is an important evolutionary force that provides abundant raw materials for genetic novelty, morphological diversity and speciation (Qiao et al., 2018). In this study, we find no evidence that *A. indica* experienced WGD after the ancient γ event shared by all eudicots. However, recent gene duplication events mixed with small-scale segmental duplication likely affected multiple genes in *A. indica*. This may also explain the fact that *A. indica* had more expanded gene families than *C. sinensis*. Our result is in agreement with the research of Chinese mahogany, which indicated that a recent WGD occurred in *Toona sinensis* (Ji et al., 2021). The occurrence of recent WGD mixed with gene duplications has been reported in *Papaver somniferum* L. genome (Guo et al., 2018). Furthermore, recent WGD was also observed in *Panax notoginseng* genome (Jiang et al., 2021). All these results are highly benefit for in-depth investigation of the survival and diversification history the of Meliaceae family.

Limonoids are natural triterpenoid products made by plants of the Meliaceae family. They are known for their insecticidal activity and potential pharmaceutical properties. *A. indica* is known as the reservoir of azadirachtin, the most famous limonoid insecticide. Secondary metabolite analysis revealed that *A. indica* contained more terpene-related clusters than that of the other 11 species. Eighty three (out of 247) clustered terpene-related genes were located on chromosome 13. The KEGG pathway enrichment analysis revealed that 33 genes were correlated with the "Sesquiterpenoid and triterpenoid biosynthesis" pathway. These results indicated that chromosome 13 may have played a central role in the evolution of terpenoid biosynthetic machinery in *A. indica*.

The TPS and CYP gene families are responsible for the biosynthesis of terpenoids in plants. 70 TPS genes were identified

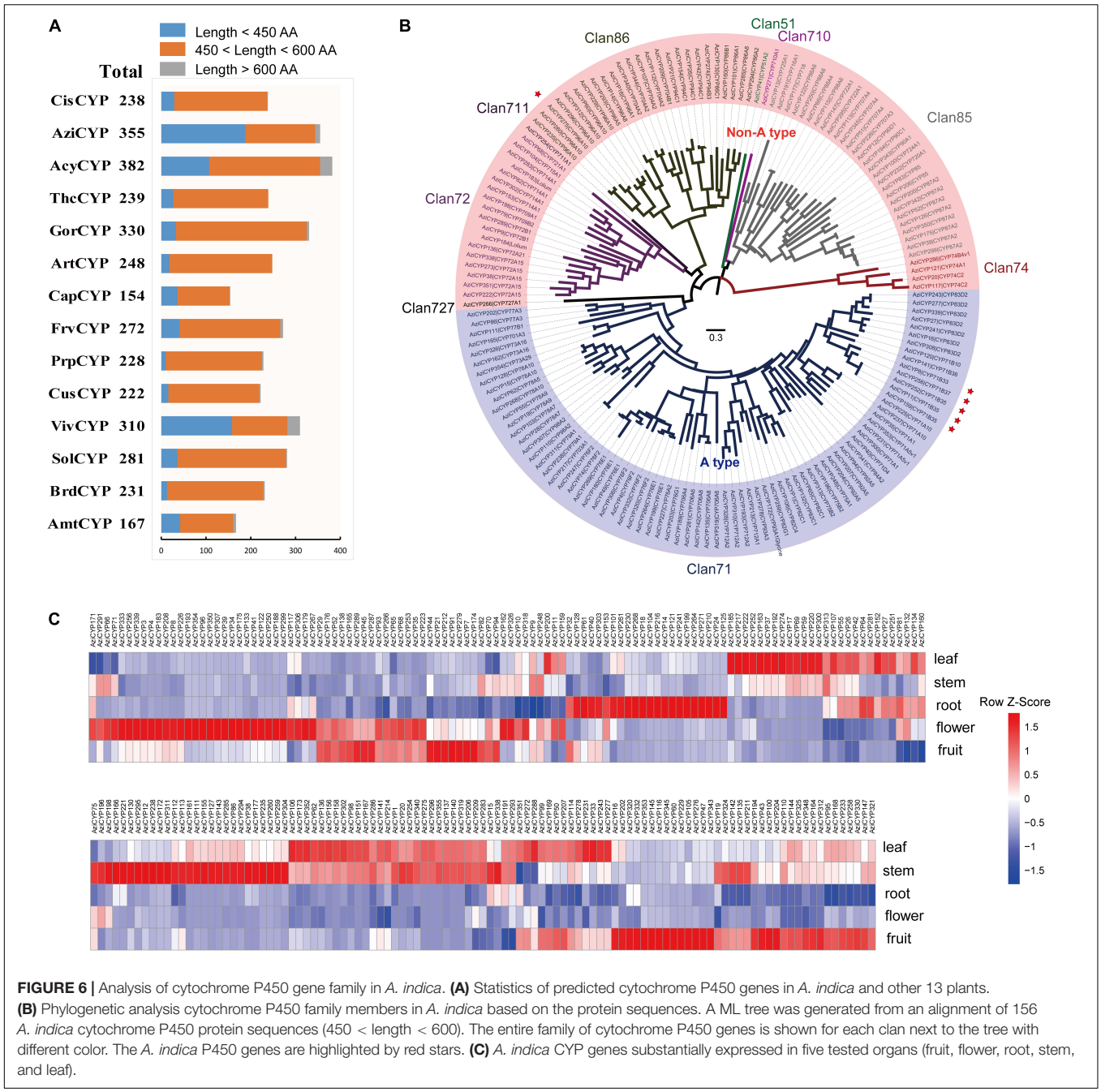

FIGURE 6 | Analysis of cytochrome P450 gene family in *A. indica*. **(A)** Statistics of predicted cytochrome P450 genes in *A. indica* and other 13 plants. **(B)** Phylogenetic analysis cytochrome P450 family members in *A. indica* based on the protein sequences. A ML tree was generated from an alignment of 156 *A. indica* cytochrome P450 protein sequences (450 < length < 600). The entire family of cytochrome P450 genes is shown for each clan next to the tree with different color. The *A. indica* P450 genes are highlighted by red stars. **(C)** *A. indica* CYP genes substantially expressed in five tested organs (fruit, flower, root, stem, and leaf).

in *A. indica*, which is much more than that of the other 13 species. This is consistent with the result of Chinese mahogany (*T. sinensis*), the first chromosome-level genome assembly of the Meliaceae family (Ji et al., 2021). Furthermore, TPS genes have also been reported to be abundant in other angiosperms that are rich in terpenoids. For example, the *Nymphaea colorata* genome harbored 92 putative TPS genes, mainly consisting of copies from subfamily TPS-b, with no TPS-a copies (Zhang et al., 2020). In contrast, more than a dozen TPS-a genes were identified in the *A. indica* genome. These TPS-a genes might be responsible for sesquiterpene biosynthesis in *A. indica*. In addition, 355 CYP genes were identified in *A. indica*, six of which were *A. indica* specific CYPs. The expansion of terpene-related gene clusters, TPSs and CYPs, may promote the formation of terpenoids in *A. indica*. A total of eight TPS genes and 36 CYP genes were involved in recent gene duplication, suggesting that recent gene duplication event may have been responsible for terpenoid biosynthesis-related gene expansion in *A. indica*, after its split from *C. sinensis*. Notably, most of the identified *A. indica* -specific TPSs and CYPs were located in the terpene-related clusters on chromosome 13, indicating that these regions were likely to be involved in azadirachtin biosynthesis. This study provided the

first chromosome-level genome of *A. indica*, and a genomic perspective for the synthesis and evolution of azadirachtin.

DATA AVAILABILITY STATEMENT

Raw data from this study were deposited in the NCBI SRA (Sequence Read Archive) database under the Bioproject ID: PRJNA645650. The genome sequence data (Illumina, PacBio, and Hi-C data) are available under accession numbers SRR12315383, SRR12321691, and SRR12321285. The assembled genome was submitted to DDBJ/ENA/GenBank with accession number JAGQDM000000000.

AUTHOR CONTRIBUTIONS

YD and WS designed the project and wrote the draft manuscript. WS participated in the genome assembly and annotation. YD, SW, JL, and ZY contributed to the genome evolution analysis, gene family analysis, and resistance gene identification. NW, HJ, JQ, and Y-XH revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This study was funded by the National Key R&D Program of China (2017YFD0201400), the Fundamental Research Funds for the Central Universities, and the General Program of National Natural Science Foundation of China (31970622).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.853861/full#supplementary-material

REFERENCES

Abdelhady, M. I. S., Bader, A., Shaheen, U., El-Malah, Y., and Barghash, M. F. (2015). Azadirachta indica as a source for antioxidant and cytotoxic polyphenolic compounds. *Biosci. Biotechnol. Res. Asia* 12, 1209–1222. doi: 10.13005/bbra/1774

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Argout, X., Salse, J., Aury, J. M., Guiltinan, M. J., Droc, G., Gouzy, J., et al. (2011). The genome of *Theobroma cacao*. *Nat. Gene.* 43, 101–108. doi: 10.1038/ng.736

Birney, E., Clamp, M., and Durbin, R. (2004). Genewise and genomewise. *Gen. Res.* 14, 988–995. doi: 10.1101/gr.1865504

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

Chen, F., Tholl, D., Bohlmann, J., and Pichersky, E. (2011). The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *Plant J.* 66, 212–229. doi: 10.1111/j.1365-313X.2011.04520.x

Chen, J. D., Zheng, C., Ma, J. Q., Jiang, C. K., Ercisli, S., Yao, M. Z., et al. (2020). The chromosome-scale genome reveals the evolution and diversification after the recent tetraploidization event in tea plant. *Hortic. Res.* 7:11. doi: 10.1038/s41438-020-0288-2

Dai, J. M., Yaylayan, V. A., Raghavan, G. S. V., Pare, J. R., and Liu, Z. (2001). Multivariate calibration for the determination of total azadirachtin-related limonoids and simple terpenoids in neem extracts using vanillin assay. *J. Agric. Food Chem.* 49, 1169–1174. doi: 10.1021/jf001141n

De Bie, T., Cristianini, N., Demuth, J. P., and Hahn, M. W. (2006). CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 22, 1269–1271. doi: 10.1093/bioinformatics/btl097

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

Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinform.* 9:14. doi: 10.1186/1471-2105-9-18

Emms, D. M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20:238. doi: 10.1186/s13059-019-1832-y

Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S. R., and Bateman, A. (2005). Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* 33, D121–D124. doi: 10.1093/nar/gki081

Guindon, S., Delsuc, F., Dufayard, J. F., and Gascuel, O. (2009). Estimating maximum likelihood phylogenies with PhyML. *Methods Mol. Biol.* 537, 113–137. doi: 10.1007/978-1-59745-251-9_6

Guo, L., Winzer, T., Yang, X., Li, Y., Ning, Z., He, Z., et al. (2018). The opium poppy genome and morphinan production. *Science* 362, 343–347. doi: 10.1126/science.aat4096

Hodgson, H., De La Pena, R., Stephenson, M. J., Thimmappa, R., Vincent, J. L., Sattely, E. S., et al. (2019). Identification of key enzymes responsible for protolimonoid biosynthesis in plants: opening the door to azadirachtin production. *Proc. Natl. Acad. Sci. U.S.A* 116, 17096–17104. doi: 10.1073/pnas.1906083116

Holt, C., and Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinform.* 12:14. doi: 10.1186/1471-2105-12-491

Hu, L., Xu, Z., Wang, M., Fan, R., Yuan, D., Wu, B., et al. (2019). The chromosome-scale reference genome of black pepper provides insight into piperine biosynthesis. *Nat. Commun.* 10:4702. doi: 10.1038/s41467-019-12607-6

Ji, Y. T., Xiu, Z., Chen, C. H., Wang, Y., Yang, J. X., Sui, J. J., et al. (2021). Long read sequencing of *Toona sinensis* (a. juss) roem: a chromosome-level reference genome for the family meliaceae. *Mol. Ecol. Resour.* 21, 1243–1255. doi: 10.1111/1755-0998.13318

Jiang, Z., Tu, L., Yang, W., Zhang, Y., Hu, T., Ma, B., et al. (2021). The chromosome-level reference genome assembly for *Panax notoginseng* and insights into ginsenoside biosynthesis. *Plant Commun.* 2:100113. doi: 10.1016/j.xplc.2020.100113

Kautsar, S. A., Suarez Duran, H. G., Blin, K., Osbourn, A., and Medema, M. H. (2017). plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Res.* 45, W55–W63. doi: 10.1093/nar/gkx305

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi: 10.1038/nmeth.3317

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly *via* adaptive k-mer weighting and repeat separation. *Genome Res.* 27, 722–736. doi: 10.1101/gr.215087.116

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinform.* 5:59. doi: 10.1186/1471-2105-5-59

Krishnan, N. M., Jain, P., Gupta, S., Hariharan, A. K., and Panda, B. (2016). An improved genome assembly of *Azadirachta indica* a. juss. *G3 (Bethesda)* 6, 1835–1840. doi: 10.1534/g3.116.030056

Krishnan, N. M., Pattnaik, S., Jain, P., Gaur, P., Choudhary, R., Vaidyanathan, S., et al. (2012). A draft of the genome and four transcriptomes of a medicinal and

pesticidal angiosperm *Azadirachta indica*. *BMC Genomics* 13:13. doi: 10.1186/1471-2164-13-464

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

Kumar, S., Stecher, G., Suleski, M., and Hedges, S. B. (2017). Timetree: a resource for timelines. timetrees, and divergence times. *Mol. Biol. Evol.* 34, 1812–1819. doi: 10.1093/molbev/msx116

Kumar, Y., Khan, F., Rastogi, S., and Shasany, A. K. (2018b). Genome-wide detection of terpene synthase genes in holy basil (*Ocimum sanctum L.*). *PLoS One* 13:e0207097. doi: 10.1371/journal.pone.0207097

Kuravadi, N. A., Yenagi, V., Rangiah, K., Mahesh, H. B., Rajamani, A., Shirke, M. D., et al. (2015). Comprehensive analyses of genomes, transcriptomes and metabolites of neem tree. *Peerj* 3:25. doi: 10.7717/peerj.1066

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

Ley, S. V. (1994). Synthesis and chemistry of the insect antifeedant azadirachtin. *Pure Appl. Chem.* 66, 2099–2102. doi: 10.1351/pac199466102099

Liu, X., Cheng, J., Zhang, G., Ding, W., Duan, L., Yang, J., et al. (2018). Engineering yeast for the production of breviscapine by genomic analysis and synthetic biology approaches. *Nat. Commun.* 9:448. doi: 10.1038/s41467-018-02883-z

Liu, Z., Suarez Duran, H. G., Harnvanichvech, Y., Stephenson, M. J., Schranz, M. E., Nelson, D., et al. (2020). Drivers of metabolic diversification: how dynamic genomic neighbourhoods generate new biosynthetic pathways in the brassicaceae. *New Phytol.* 227, 1109–1123. doi: 10.1111/nph.16338

Lowe, T. M., and Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25, 955–964. doi: 10.1093/nar/25.5.955

Mithilesh, S., and Rakhi, C. (2014). Sustainable production of azadirachtin from differentiated *in vitro* cell lines of neem. *AoB Plants* 5, lt034–lt034.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., and Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* 35, W182–W185. doi: 10.1093/nar/gkm321

Nawrocki, E. P., and Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinform.* 29, 2933–2935. doi: 10.1093/bioinformatics/btt509

Nowak, M. D., Russo, G., Schlapbach, R., Huu, C. N., Lenhard, M., and Conti, E. (2015). The draft genome of *Primula veris* yields insights into the molecular basis of heterostyly. *Genome Biol.* 16:16. doi: 10.1186/s13059-014-0567-z

Nutzmann, H. W., Huang, A., and Osbourn, A. (2016). Plant metabolic clusters - from genetics to genomics. *New Phytol.* 211, 771–789. doi: 10.1111/nph.13981

Osuna-Cruz, C. M., Paytuvi-Gallart, A., Di Donato, A., Sundesha, V., Andolfo, G., Cigliano, R. A., et al. (2018). PRGdb 3.0: a comprehensive platform for prediction and analysis of plant disease resistance genes. *Nucleic Acids Res.* 46, D1197–D1201. doi: 10.1093/nar/gkx1119

Ou, S., Chen, J., and Jiang, N. (2018). Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* 46:e126. doi: 10.1093/nar/gky730

Ou, S. J., and Jiang, N. (2018). LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* 176, 1410–1422. doi: 10.1104/pp.17.01310

Pellicer, J., and Leitch, I. J. (2020). The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. *New Phytol.* 226, 301–305. doi: 10.1111/nph.16261

Powell, S., Szklarczyk, D., Trachana, K., Roth, A., Kuhn, M., Muller, J., et al. (2012). eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.* 40, D284–D289. doi: 10.1093/nar/gkr1060

Price, A. L., Jones, N. C., and Pevzner, P. A. (2005). De novo identification of repeat families in large genomes. *Bioinform.* 21, I351–I358. doi: 10.1093/bioinformatics/bti1018

Qiao, X., Yin, H., Li, L., Wang, R., Wu, J., Wu, J., et al. (2018). Different modes of gene duplication show divergent evolutionary patterns and contribute differently to the expansion of gene families involved in important fruit traits in pear (*Pyrus bretschneideri*). *Front. Plant Sci.* 9:161. doi: 10.3389/fpls.2018.00161

Rodrigues, M., Festucci-Buselli, R. A., Silva, L. C., and Otoni, W. C. (2014). Azadirachtin biosynthesis induction in *Azadirachta indica* a. juss cotyledonary calli with elicitor agents. *Braz. Arch. Biol. Technol.* 57, 155–162. doi: 10.1590/s1516-89132014000200001

Schmutterer, H. (1995). The neem tree, *Azadirachta indica* a. juss. and other meliaceous plants: source of unique natural products for integrated pest management, medicine, industry and other purposes. *Pap. Bibliogr. Soc. Am.* 107, 1365–1372.

Servant, N., Varoquaux, N., Lajoie, B. R., Viara, E., Chen, C. J., Vert, J. P., et al. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* 16:11. doi: 10.1186/s13059-015-0831-x

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351

Soares, D., Godin, A., Menezes, R., Nogueira, R., Brito, A., Melo, I., et al. (2014). Anti-inflammatory and antinociceptive activities of azadirachtin in mice. *Planta Med.* 80, 630–636. doi: 10.1055/s-0034-1368507

Srivastava, S., and Srivastava, A. K. (2013). Production of the biopesticide azadirachtin by hairy root cultivation of azadirachta indica in liquid-phase bioreactors. *Appl. Biochem. Biotechnol.* 171, 1351–1361. doi: 10.1007/s12010-013-0432-7

Stanke, M., Steinkamp, R., Waack, S., and Morgenstern, B. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* 32, W309–W312. doi: 10.1093/nar/gkh379

Stix, G. (1992). Village pharmacy. the neem tree yields products from pesticides to soap. *Sci. Am.* 266:132. doi: 10.1038/scientificamerican0592-132

Tang, H., Bowers, J. E., Wang, X., Ming, R., Alam, M., and Paterson, A. H. (2008). Synteny and collinearity in plant genomes. *Science* 320, 486–488. doi: 10.1126/science.1153917

Tarailo-Graovac, M., and Chen, N. (2009). Using repeatmasker to identify repetitive elements in genomic sequences. *Curr. protoc. Bioinform.* Chapter 4, Unit 4.10. doi: 10.1002/0471250953.bi0410s25

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with tophat and cufflinks. *Nat. Protoc.* 7, 562–578. doi: 10.1038/nprot.2012.016

Tu, L., Su, P., Zhang, Z., Gao, L., Wang, J., Hu, T., et al. (2020). Genome of Tripterygium wilfordii and identification of cytochrome P450 involved in triptolide biosynthesis. *Nat. Commun.* 11, 971. doi: 10.1038/s41467-020-14776-1

Vaser, R., Sovic, I., Nagarajan, N., and Sikic, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* 27, 737–746. doi: 10.1101/gr.214270.116

Veitch, G. E., Beckmann, E., Burke, B. J., Boyer, A., Maslen, S. L., and Ley, S. V. (2007). Synthesis of azadirachtin: a long but successful journey. *Angew. Chem. Int. Ed Engl.* 46, 7629–7632. doi: 10.1002/anie.200703027

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9:14. doi: 10.1371/journal.pone.0112963

Wang, D., Zhang, Y., Zhang, Z., Zhu, J., and Yu, J. (2010). KaKs_calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinform.* 8, 77–80. doi: 10.1016/s1672-0229(10)60008-3

Wang, H., Wang, N., and Huo, Y. (2020). Multi-tissue transcriptome analysis using hybrid-sequencing reveals potential genes and biological pathways associated with azadirachtin a biosynthesis in neem (*azadirachta indica*). *BMC Genomics* 21:749. doi: 10.1186/s12864-020-07124-6

Xu, Q., Chen, L. L., Ruan, X. A., Chen, D. J., Zhu, A. D., Chen, C. L., et al. (2013). The draft genome of sweet orange (Citrus sinensis). *Nat. Genetics* 45, 59–U92. doi: 10.1038/ng.2472

Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* 35, W265–W268. doi: 10.1093/nar/gkm286

Yang, J., Wariss, H. M., Tao, L. D., Zhang, R. G., Yun, Q. Z., Hollingsworth, P., et al. (2019). De novo genome assembly of the endangered *Acer yangbiense*, a plant species with extremely small populations endemic to yunnan province. *China. Gigascience* 8:10. doi: 10.1093/gigascience/giz085

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* 13, 555–556. doi: 10.1093/bioinformatics/13.5.555

Zhang, L., Chen, F., Zhang, X., Li, Z., Zhao, Y., Lohaus, R., et al. (2020). The water lily genome and the early evolution of flowering plants. *Nature* 577, 79–84. doi: 10.1038/s41586-019-1852-5

Zhang, X. T., Zhang, S. C., Zhao, Q., Ming, R., and Tang, H. B. (2019). Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* 5, 833–845. doi: 10.1038/s41477-019-0487-8

Zhang, Z., Xiao, J., Wu, J., Zhang, H., Liu, G., Wang, X., et al. (2012). ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments. *Biochem. Biophys. Res. Commun.* 419, 779–781. doi: 10.1016/j.bbrc.2012.02.101

Zhou, H. C., Shamala, L. F., Yi, X. K., Yan, Z., and Wei, S. (2020). Analysis of terpene synthase family genes in *Camellia sinensis* with an emphasis on abiotic stress conditions. *Sci. Rep.* 10:933. doi: 10.1038/s41598-020-57805-1

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

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

Copyright © 2022 Du, Song, Yin, Wu, Liu, Wang, Jin, Qiao and Huo. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 12 May 2022
doi: 10.3389/fpls.2022.882960

An NGS-Based Phylogeny of Orthotricheae (Orthotrichaceae, Bryophyta) With the Proposal of the New Genus *Rehubryum* From Zealandia

Isabel Draper[1,2], Tamara Villaverde[3,4], Ricardo Garilleti[5], J. Gordon Burleigh[6], Stuart F. McDaniel[6], Vicente Mazimpaka[1,2], Juan A. Calleja[1,2] and Francisco Lara[1,2]*

[1]Centro de Investigación en Biodiversidad y Cambio Global, Universidad Autónoma de Madrid, Madrid, Spain, [2]Departamento de Biología, Facultad de Ciencias, Universidad Autónoma de Madrid, Madrid, Spain, [3]Departamento de Biodiversidad, Ecología y Evolución,Universidad Complutense de Madrid, Madrid, Spain, [4]Departamento de Biología, Geología, Física y Química Inorgánica, Universidad Rey Juan Carlos, Móstoles, Spain, [5]Departamento de Botánica y Geología, Facultad de Farmacia, Universidad de Valencia, Valencia, Spain, [6]Department of Biology, University of Florida, Gainesville, FL, United States

OPEN ACCESS

Edited by:
Gerald Matthias Schneeweiss, University of Vienna, Austria

Reviewed by:
Alain Vanderpoorten, University of Liège, Belgium
Rafael Hernández Maqueda, University of Almeria, Spain

***Correspondence:**
Isabel Draper
isabel.draper@uam.es

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

Received: *24 February 2022*
Accepted: *25 April 2022*
Published: *12 May 2022*

Citation:
Draper I, Villaverde T, Garilleti R, Burleigh JG, McDaniel SF, Mazimpaka V, Calleja JA and Lara F (2022) An NGS-Based Phylogeny of Orthotricheae (Orthotrichaceae, Bryophyta) With the Proposal of the New Genus Rehubryum From Zealandia. Front. Plant Sci. 13:882960. doi: 10.3389/fpls.2022.882960

Phylogenomic data increase the possibilities of resolving the evolutionary and systematic relationships among taxa. This is especially valuable in groups with few and homoplasious morphological characters, in which systematic and taxonomical delimitations have been traditionally difficult. Such is the case of several lineages within Bryophyta, like Orthotrichaceae, the second most diverse family of mosses. Members of tribe Orthotricheae are common in temperate and cold regions, as well as in high tropical mountains. In extratropical areas, they represent one of the main components of epiphytic communities, both in dry and oceanic or hyperoceanic conditions. The epiphytic environment is considered a hostile one for plant development, mainly due to its low capacity of moisture retention. Thus, the diversification of the Orthotrichaceae in this environment could be seen as striking. Over the last two decades, great taxonomic and systematic progresses have led to a rearrangement at the generic level in this tribe, providing a new framework to link environment to patterns of diversification. Here, we use nuclear loci targeted with the GoFlag 408 enrichment probe set to generate a well-sampled phylogeny with well-supported suprageneric taxa and increasing the phylogenetic resolution within the two recognized subtribes. Specifically, we show that several genera with *Ulota*-like morphology jointly constitute an independent lineage. Within this lineage, the recently described *Atlantichella* from Macaronesia and Western Europe appears as the sister group of *Ulota bellii* from Zealandia. This latter species is here segregated in the new genus *Rehubryum*. Assessment of the ecological and biogeographical affinities of the species within the phylogenetic framework suggests that niche adaptation (including climate and substrate) may be a key evolutionary driver that shaped the high diversification of Orthotricheae.

Keywords: Orthotrichinae, Lewinskyinae, *Ulota bellii*, *Atlantichella*, *Plenogemma*, *Pulvigera*, GoFlag 408 hyb seq

INTRODUCTION

During the last decades, the use of molecular phylogenetics has transformed our understanding of biodiversity. Molecular data have provided a powerful tool both to resolve phylogenetic relationships and to delimitate the boundaries of taxa at different taxonomic levels. Whereas much progress has been made using data from Sanger sequencing, Next-Generation Sequencing (NGS) techniques can generate far more data and thus greatly enhance the resolution of the genealogy of life (e.g., Lemmon and Lemmon, 2013; Weitemier et al., 2014; Villaverde et al., 2018; Shah et al., 2021).

One of the most promising NGS methods is target enrichment (Weitemier et al., 2014), which can generate data from hundreds, if not thousands, of low-copy nuclear loci that can be used to reconstruct plant phylogenies (e.g., Liu et al., 2019). Recently, the GoFlag project developed a target enrichment probe set that can be used all along the flagellate land plants (i.e., bryophytes, lycophytes, ferns, and gymnosperms; Breinholt et al., 2021). Data generated from these kits can help resolve backbone relationships within large phylogenies and help elucidate important evolutionary processes such as the diversification of land plants. In addition, they also provide an enormous amount of genetic information that can help to resolve the relationships among closely related species or even among populations (e.g., Villaverde et al., 2018). This is especially valuable in groups with few, and often homoplasious, morphological characters, in which systematic and taxonomic delimitations, have been traditionally difficult and contentious.

Bryophytes, including mosses, liverworts, and hornworts, have been repeatedly considered to be genetically static. For example, Liu et al. (2014) provided evidence of the conservation and stasis of the mitochondrial genome in mosses for over 350 My. Similarly, Rosato et al. (2016) postulated evolutionary rDNA stasis during land colonization and diversification across 480 My of bryophyte evolution, and Dong and Liu (2021) stated that genome structure is also static, especially in mosses. This genetic stability correlates, at least in some cases, with a morphological stasis, as demonstrated by McDaniel and Shaw (2003) for *Pyrrhobryum mnioides* (Hook.) Manuel or Aigoin et al. (2009) for *Hedenasiastrum* Ignatov & Vanderp. Nevertheless, the lack of morphological characters can also be due to recent speciation, since relatively young sister species may have had insufficient time to develop and accumulate phenotypic differences (Renner, 2020). This is especially true in bryophytes, where the low structural complexity of the dominant gametophyte implies fewer taxonomically relevant morphological characters in comparison with other groups of land plants. In such groups, the use of NGS techniques may be especially useful for understanding their relationships.

The mosses are the most diverse bryophyte lineage (Liu et al., 2019) with *ca*. 120 families, and Orthotrichaceae Arn., with an estimated 850 species, is the second most diverse family (Frey and Stech, 2009). Orthotrichaceae comprises two subfamilies, Orthotrichoideae Broth. and Macromitrioideae Broth., which differ in morphological, biogeographic, and ecological traits (Lara et al., 2014). Macromitrioideae is almost exclusively intertropical, whereas Orthotrichoideae is common in temperate and cold regions of both hemispheres, as well as in high tropical mountains. Orthotrichoideae is, in turn, divided into two tribes, Orthotricheae Engler and Zygodonteae Engler (Goffinet and Vitt, 1998; Draper et al., 2021). Of these, Orthotricheae stands out as one of the main components of the epiphytic communities in temperate areas, both in dry (Draper et al., 2006; Lara et al., 2009) and in oceanic or hyperoceanic conditions (Garilleti et al., 2015; Lara et al., 2016). Species within Orthotricheae are acrocarpous mosses, whose gametophores grow erect or rarely decumbent and form cushions or tufts. Their leaves are variously lanceolate, erect or imbricate, sometimes twisted when dry, with upper cells rounded, papillose, and basal cells enlarged, smooth, and with single nerves ending near the leaf apex. Their calyptrae are mitrate, plicate, and commonly hairy, and their sporophytes are either immersed, emergent, or variously exserted, with capsules mostly cylindric, smooth, or commonly furrowed, bearing superficial or immersed stomata, and a double peristome of 16 exostomial teeth alternating with 16 endostomial segments that could be somewhat modified or variably reduced (see, e.g., Lara et al., 2014). In fact, the gametophytes of the different Orthotricheae species are overall morphologically similar, and few species within this group can be identified in the absence of sporophytes. This may be one of the reasons for the numerous taxonomic changes that this group has experienced during the last decades, involving rearrangements affecting the main genera (Goffinet et al., 2004; Plášek et al., 2015; Lara et al., 2016; Draper et al., 2021).

For many years, Orthotricheae was understood to include only two large genera, *Orthotrichum* Hedw. and *Ulota* D.Mohr, but in the last 20 years, these two genera have been, respectively, split into five and three genera (see a revision in Draper et al., 2021). Two phylogenetic reconstructions including a representative selection of the Orthotricheae taxa as currently understood have been published recently (Draper et al., 2021; Wang et al., 2021), based on the analyses of 4 nuclear and chloroplast loci, and 6 loci from all three plant genomes, respectively. These two studies represent an important step forward for the understanding of the possible evolutionary history of this tribe. However, the relationships among some taxa remained unresolved, especially within subtribe Lewinskyinae F.Lara, Garilleti & Draper. Also, both of these studies are based on a limited number of loci and could potentially be misled due to complex evolutionary processes such as incomplete lineage sorting or reticulate evolution that are difficult to resolve with few loci (see, e.g., Degnan and Rosenberg, 2009).

Few attempts have been made to reconstruct the relationships in this group with larger phylogenomic data. In addition to the 6 loci previously mentioned, Wang et al. (2021) analyzed 40 mitochondrial and 82 chloroplast genes for a subset selection of 23 Orthotricheae taxa. These authors provided additional evidence supporting the currently accepted circumscription of genera, but the genus *Atlantichella* F.Lara, Garilleti & Draper was not included. In addition, their sampling included few species from each genus. This sparse sampling is especially notable in the largest three genera: *Orthotrichum* (6 out of

~100 species), *Ulota* (4 out of 70 species), and *Lewinskya* F.Lara, Garilleti & Goffinet (6 out of 70 species). Due to the limited sampling, this study could not address the relationships within Lewinskyinae or discern biogeographic and evolutionary patterns that may help explain the diversification and distribution of Orthotricheae.

In this study, we aim to expand upon recent phylogenetic studies with the analysis of a large-scale nuclear dataset generated using target enrichment and including a wide representation of Orthotricheae species, including the main genera and their subgenera, as traditionally delimited. Specifically, we intend to answer whether the nuclear data support the current delimitation of Orthotricheae at the genus level. As indicated above, this has been tested through the analyses of organellar genomes, but nuclear data have been restricted to the inclusion of *ITS*2 and 26*S* in the phylogenies by Draper et al. (2021) and Wang et al. (2021), respectively. The epiphytic habitat is characterized by its low moisture retention capacity, which is especially harsh in areas with climates including a dry season (e.g., Pugnaire and Valladares, 1999). Consequently, the high diversification of *Orthotrichum*, *Ulota,* and *Lewinskya* in this environment could be seen as striking. As a first step toward explaining the high diversification of Orthotricheae in the hostile epiphytic environment, we assess ecological and biogeographic affinities within the herein generated phylogenetic framework to identify putative drivers of evolutionary success in this group.

MATERIALS AND METHODS

Taxon Sampling

We sampled 80 taxa of Orthotrichoideae, focusing on tribe Orthotricheae, with representatives of 7 of its 9 genera. Details on the samples included in the analyses are shown in **Supplementary Table S1**, with nomenclature following Tropicos (2022) database and abbreviations of authors of plant names following IPNI (2021) database. Specifically, we included 72 species of Orthotricheae, which constitutes approximately 30% of the accepted species: 26 species (out of 100 accepted) of *Orthotrichum*, 1 (of 2) of *Nyholmiella* Holmen & E.Warncke, 22 (of 70) of *Lewinskya*, 1 (of 4) of *Pulvigera* Plášek, Sawicki & Ochyra, 20 (of 70) of *Ulota*, 1 (of 1) of *Plenogemma* Plášek, Sawicki & Ochyra, and 1 (of 1) of *Atlantichella*.

To root the phylogenetic tree, we used representatives of tribe Zygodonteae as sister group, with 4 species of *Zygodon* Hook. & Taylor and 1 of *Australoria* F.Lara, Garilleti & Draper. As ultimate outgroup, we included 1 species of *Leratia* Broth. & Paris (of the sister subfamily Macromitrioideae).

DNA Extraction

We extracted DNA from the selected samples with a modified cetyltrimethylammonium bromide (CTAB) protocol (Doyle and Doyle, 1987) described in Breinholt et al. (2021). We used a Geno/Grinder 2010 mill (SPEX CertiPrep, Metuchen, New Jersey, United States) to lyse the cells and performed two rounds of chloroform washes followed by an isopropanol precipitation and an ethanol wash. We added 2 µl of 10 mg/ml RNase A (QIAGEN, Valencia, California, United States) to each sample between chloroform washes to remove RNA contamination.

Target Enrichment and Sequencing Assembly

We employed a target enrichment approach using the GoFlag 408 probe set to generate a multi-locus nuclear sequence dataset for phylogenetic analyses. The GoFlag 408 probe set targets 408 exons found in 229 single or low-copy nuclear genes and appears to recover many loci across mosses (Breinholt et al., 2021). Library preparation, target enrichment, and sequencing were done by RAPiD Genomics (Gainesville, FL, USA). Protocols for library preparation and hybridization are described in Breinholt et al. (2021). All enriched, pooled libraries were sequenced on an Illumina HiSeq 3,000 platform (Illumina; 2 × 100 bp).

We extracted the targeted loci from the raw sequence reads using a pipeline described in detail in Breinholt et al. (2021), and the scripts and reference sequences are available in Dryad (Breinholt et al., 2020). We trimmed the raw sequence reads with Trim Galore! Version 0.4.4[1] to remove adapters and bases with a Phred score below 20. We then assembled the targeted loci for each sample using iterative baited assembly (IBA; Breinholt et al., 2018), which conducts a *de novo* assembly with BRIDGER version 2014-12-01 (Chang et al., 2015) based on sequence homology of raw reads to a set of reference sequences for each locus. The IBA seeks to extend the assemblies beyond the target regions and recover as much of the more variable flanking intron regions as possible. In this sense, two types of matrices were generated as: (a) assembled sequences trimmed to the probe region (referred to as Probe Only matrices) and (b) full-length assembled sequences, i.e., including probe regions as well as flaking intron sequences (referred as Full Sequences matrices).

Next, we performed an orthology assessment using the target region sequences based on a tBLASTx (Camacho et al., 2009) search against nine flagellate plant genomes to remove potential paralogs. We also removed possible contaminants by performing a tBLASTx search of the assemblies for each locus against the reference sequences, and we removed any sequences that had the best hit that did not come from a moss. Finally, for each locus, we aligned the recovered sequences from the target regions only (Probe Only) and from the combined target and flanking regions from each locus (Full Sequences) with MAFFT 7.425 (Katoh and Standley, 2013). Putative isoforms from the same taxon were merged with a Perl script that used IUPAC ambiguity codes to represent putative heterozygous sites.

Although this pipeline does not explicitly phase loci, in some cases, a locus alignment might include multiple sequences from some samples, representing cases in which the BRIDGER assembler identified allelic diversity. In these cases, to reduce the possibility of including paralogs in our phylogenetic analyses, for each sample with more than one sequence, we removed

[1]https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

all sequences from that sample from the locus alignment. Across the 405 loci from which we recovered sequences from at least four samples, we removed 1934 sequences, while retaining 25,544 single-copy sequences. Also, alignments of the flanking regions often have large gaps with sequences from one or a few samples due to indels and the high variability in the recovered length of the flanking sequence. Thus, to clean the alignment and reduce missing data, we ran a script to remove all columns from the locus alignments with fewer than four nucleotides. Additionally, the Probe Only matrix was also pruned with Gblocks (Castresana, 2000) with the following settings: -t = d -b1 = 51 -b2 = 60 -b3 = 8 -b4 = 8 -b5 = h. Summary statistics were calculated using AMAS (Borowiec, 2016). Finally, individual matrices from both datasets were concatenated into two independent supermatrices (i.e., Full Sequences supermatrix and Probe Only supermatrix).

Phylogenetic Inference

Trees were inferred using two approaches: (a) a total evidence approach using maximum likelihood (ML) inference based on a concatenated matrix of all loci and (b) a summary species tree method that accounts for the Multiple Species Coalescent (MSC) with ASTRAL-III 5.7.8 (Zhang et al., 2018).

For the total evidence approach, we ran ML analyses of both the Full Sequences supermatrix and the Probe Only supermatrix. Phylogenetic analyses of these two datasets were executed in IQ-TREE 2.0.3 (Minh et al., 2020), after automatic model selection using ModelFinder (Kalyaanamoorthy et al., 2017) with the approximate likelihood ratio test ("-alrt" option). These analyses also included 1,000 bootstrap replicates and 1,000 ultrafast bootstrap ("bb" option). To investigate gene tree versus species tree concordance, we 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. This approach provides a description of possible disagreement among loci and across sites within the sequence. We considered only branches with ultrafast-bootstrap support values >90% as statistically supported. Trees were plotted in FigTree 1.4.4.[2]

For the summary species tree approach under the MSC, individual gene trees were constructed using RAxML 8.2.12 (Stamatakis, 2014) applying GTR-CAT and 200 bootstrap replicates followed by slow ML optimization with the "-f a" option. Then, branches with BS < 50% were collapsed using Newick utilities (Junier and Zdobnov, 2010). Species tree inference under the MSC approach was then performed using ASTRAL-III, and branch support values were inferred through local posterior probabilities (LPP; Sayyari and Mirarab, 2016). Values of LPP > 0.95 were considered to represent strong branch support, although lower values (LPP = 0.7–0.9) also may indicate high support (Sayyari and Mirarab, 2016). To output quartet support values, we used the "-t 2" option. We plotted pie charts reporting the proportion of quartet values in R (R Core Team, 2020) using the packages ape (Paradis and Schliep, 2019), ggimage (Yu, 2021), ggtree (Yu et al., 2017), treeio (Yu et al., 2017), and their corresponding dependencies.

[2]https://github.com/rambaut/figtree/releases

Niche Preference Characterization

We categorized the studied species according to their niche preferences (regarding substrate and climate) and distribution. Orthotricheae mosses occur on three types of substrates: rocks, tree trunks (including large branches), and small branches (including twigs). All the species were characterized according to their preferences for one, two, or all three possible substrates, mainly on the basis of the expert knowledge of the authors and always recording the prevailing ecological behavior of the species, without considering the most exceptional situations (Mazimpaka and Lara, 1995). We classified the climatic preferences of each species regarding the degree of humidity of the climatic environment in which the species usually grow. These were also divided into three principal types: arid (with scarce precipitations and long periods of dry season, such as the Mediterranean climate), dry (with scarce to moderate precipitations, but without long periods of dry season), or wet (humid or hyper-humid climates, including local or regional situations with frequent mists that produce horizontal precipitations), and we characterized the species as showing preferences for one, two, or all three of them. Finally, we described the climatic preferences regarding the degree of thermicity according to the latitudinal bands where the species thrive as temperate (including cold-temperate), subtropical, and tropical-montane. Subtropical is used as defined by Troll and Paffen (1964) and includes warm climates, between tropical and temperate, with mean temperatures between 17 and 24°C, as prevail in the Mediterranean basin, southern California, The Cape Region, or Macaronesia. The characterization of the species was completed with a description of their geographical range: Subcosmopolite, Holarctic, Sub-antarctic, Australasia, Europe, Western Europe, Mediterranean basin, Macaronesia, East Africa, South Africa, East Asia, South India, North America, Central America, Caribbean, South America, Tropical Andes, and Patagonia.

RESULTS

Capture Success and Data Quality

The target enrichment recovered more than 200 nuclear loci (out of a possible 408) for 71 out of 80 samples, and more than 350 loci for 56 of these samples. Five samples largely failed, recovering 5 or fewer loci (*Lewinskya acuminata*, *L. sordida*, *L. tasmanica*, *O. pilosissimum,* and *O. rivulare*), and consequently, we did not include these samples in our analyses. Only five of the targeted loci were recovered in fewer than 10 samples. The average proportion of missing data in the Probe Only supermatrix was 0.63%, and only 10 out of 408 loci had more than 5% of missing data. The average proportion of parsimony informative sites per locus was 17.6%, and only 22 out of 408 loci had less than 10% of parsimony informative sites. The Full Sequences supermatrix was notably noisier, with an average proportion of missing data of 47%, and only 2 loci with less

than 5% of missing data. Summary statistics are available in **Supplementary Table S1** (sample statistics including number of loci and percentage of sequencing success) and **Supplementary Table S2** (loci statistics including length, number of taxa, and number of variable, parsimony informative and missing data sites for both the Probe Only and the Full Sequences supermatrices). Raw data files are available in the GenBank Sequence Read Archive (SRA) under the BioProject number PRJNA819401. The unique accession number of each sample is available in **Supplementary Table S3**.

Phylogenetic Reconstruction

The analyses of the Full Sequences supermatrix and the Probe Only supermatrix yielded phylogenetic trees with similar topologies, although the Full Sequences trees showed shorter branches than the Probe Only trees for some of the nodes in the in-group. Therefore, the trees shown on **Figure 1** and commented hereafter are those resulting from the Probe Only supermatrix analyses. The trees based on the Full Sequences supermatrix are included as **Supplementary Figures**, as well as pie charts reporting quartet support values for the Probe Only analyses.

Both species-based trees (IQ-TREE) and gene-based trees (ASTRAL) recover overall concordant clades (**Figure 1**). Many nodes are well supported with IQ-TREE ultrafastbootstrap (BS; i.e., values greater than 90%) and ASTRAL posterior probability (LPP; i.e., values greater than 0.9). Some of the branches that receive low BS or LPP value support have low gCF scores and/or have low quartet scores (**Supplementary Table S4**).

Samples of Orthotricheae were included in a single clade sister to samples of Zygdonteae in all the analyses. Within the Orthotricheae clade, samples were distributed in two monophyletic groups, one including the samples of *Orthotrichum* and *Nyholmiella* (not well supported clade E, dark green colored in **Figure 1**) and the other including the samples from the remaining genera (highly supported clade A, light green). Within clade E, all the analyses placed the sample of *Nyholmiella* as sister to a highly supported clade F, which included all the samples of the species of *Orthotrichum*. This *Orthotrichum* clade F was in turn divided into subclades according to both the species- and gene-based trees, although not all these inner clades were maximally supported, and there was some incongruence on the grouping of *Orthotrichum patens*, *O. stramineum*, *O. alpestre*, *O. consobrinum*, *O. cucullatum*, and *O. consimile*.

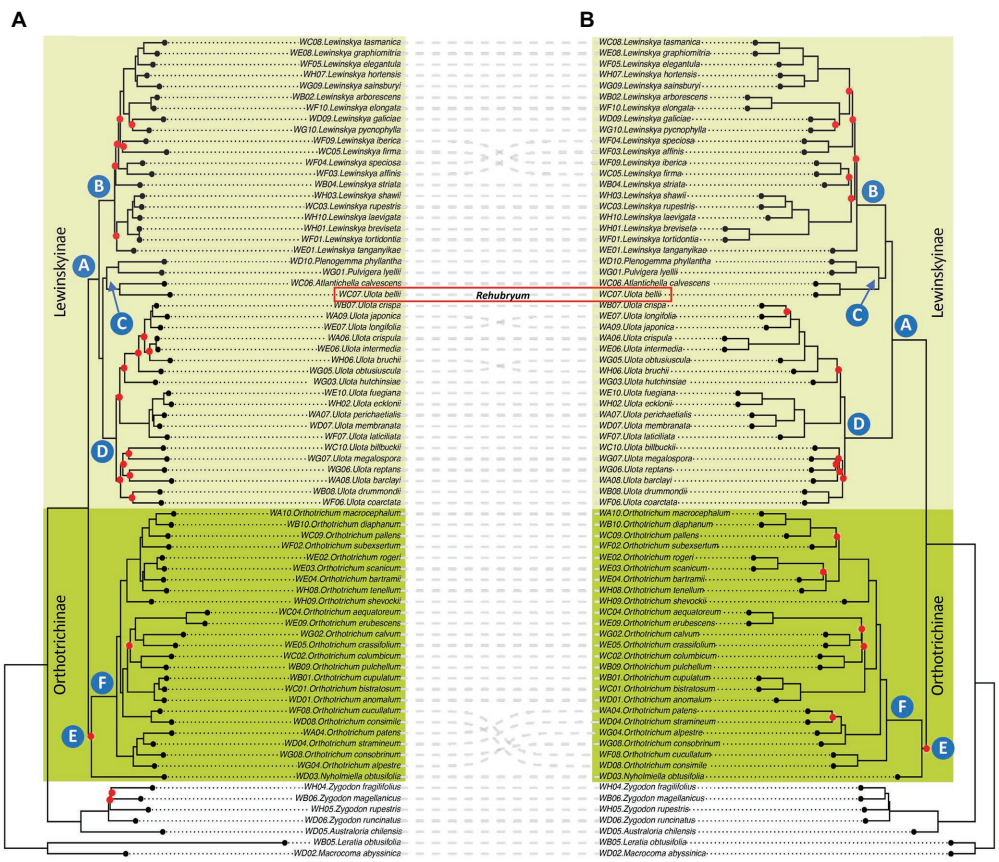

FIGURE 1 | Phylogenetic reconstructions: **(A)** IQ-TREE analysis of the Probe Only supermatrix: concatenation of 21,501 bp after Gblocks pruning. Red dots indicate nodes with ultrafast BS lower than 90. **(B)** ASTRAL analysis of the Probe Only supermatrix: consensus phylogeny based on a total of 405 trees after tBLASTx, with branches with less than 50 BS collapsed. Red dots indicate nodes with LPP lower than 0.90.

Regarding clade A, all the analyses established an inner subgrouping in three large and strongly supported clades, named B, C, and D in **Figure 1**. The relationships of these three clades varied depending on the analysis: the species-based tree supported a closer relationship between clades C and D, and placed B as sister of the two, whereas the gene-based tree supported a sister relationship for B and C, and placed D as sister of them. Clade B included all the samples of *Lewinskya* and was in turn divided into inner subclades both according to the species- and gene-based trees, although this inner grouping was not maximally supported in any of the analyses. The groups resulting from both the species- and gene-based trees were overall congruent, except for the grouping of *Lewinskya speciosa*, *L. affinis*, *L. iberica,* and *L. firma*. Similarly, clade D was divided into subclades (not fully supported) that were overall congruent among the species- and gene-based trees, except for the relationships established for *Ulota longifolia*, *U. japonica*, *U. bruchii* and *U. hutchinsiae*. This clade D included all the samples of *Ulota* with the exception of the sample of the species *U. bellii*, which was included in clade C together with the small genera *Plenogemma*, *Pulvigera,* and *Atlantichella*.

Ecological and Biogeographical Affinities

The ecological and biogeographical affinities of the studied species were plotted in the phylogenetic framework to assess whether the recovered clades could reflect ecological or biogeographical patterns. As explained above, the main clades (i.e., clades representing the genus taxonomic level) established by all the analyses performed were congruent, even though the relationships among them sometimes differed. This was the case for the position of clade C, which includes three taxa (*Plenogemma phyllantha*, *Atlantichella calvescens,* and *Ulota bellii*) that have been traditionally treated as *Ulota* due to their morphological similarities. These similitudes were supported by the IQ-TREE analysis, which showed a sister relationship of clade C and *Ulota*. For this reason, we selected the IQ-TREE phylogenetic reconstruction to plot the ecological and biogeographical affinities, in order to identify the possible patterns underlying the recovered clades (**Figure 2**).

Most of the established clades were clearly congruent with the substrate and climatic preferences. Thus, *Lewinskya* species as a whole tend to be specialized in colonizing tree trunks and large branches, although most of them often also grow on small branches and twigs. Regarding humidity, species of this genus tend to show preferences for dry climate, although some species are typical or common in wet areas, and *L. rupestris*, a cortico-saxicolous species, shows a wide range of humidity tolerance. The highest variability is found among the temperature preferences that range from cold-temperate to tropical-montane at the genus level. Notably, this variability agreed in general terms with the inner subclades of *Lewinskya*, even though these subclades were not always maximally supported. For example, the maximally supported clade including *L. tasmanica* and related species could be defined as typically temperate, whereas the not maximally supported clade of *L. arborescens* and related species could be defined as typically tropical-montane. Conversely, the clades recovered did not show a clear geographical pattern, and species that usually coexist, such as *L. breviseta* and *L. iberica* in the Mediterranean or *L. arborescens* and *L. firma* in East Africa, were not closely related.

The clade including the species of *Ulota* was also clearly differentiated from the rest by its ecological preferences (**Figure 2**). Regarding substrate affinities, this group of species shows a clear tendency toward small branches and twigs (except for the mainly saxicolous *U. hutchinsiae*), although most of them can also appear on large branches and trunks. As for climatic preferences, these species are all typical of wet areas, temperate, or cold-temperate, from both hemispheres. Exceptionally, they can also thrive in subtropical (Lara et al., 2022) or tropical-montane areas (Garilleti et al., 2015). Again, no clear biogeographical pattern could be established for this genus.

The third large genus of the family is *Orthotrichum*. As shown in **Figure 2**, this genus is more variable regarding the ecological preferences of its species than the two above mentioned, although some ecological patterns were also obtained, especially for the inner subclades. Thus, this genus includes species usually growing on rocks (some of which were grouped in the subclade including *O. anomalum* and related species), on trunks and large branches (such as those included in the subclade around *O. subexsertum*), and (more rarely) on small branches and twigs. Regarding its climatic preferences, it ranges from an affinity for wet areas to dry or even arid ones and from cold-temperate to tropical-montane areas. No clear pattern was recovered that could easily explain the phylogenetic clades recovered, either based on the climatic or biogeographical affinities.

The remaining five genera of Orthotricheae included in this analysis were represented by a single species each. Four of these were grouped into clade C that, according to the species-based reconstruction, is sister to *Ulota* (**Figures 1, 2**). This clade C did not represent any biogeographical pattern, since, e.g., *Atlantichella calvescens* and *Ulota bellii*, inferred to be sister taxa in all the analyses, show an antipodal distribution. Conversely, they share a clear preference for small branches and twigs, and climatic preferences for wet areas, from cold-temperate extending to the subtropical latitudes in the case of *A. calvescens*.

DISCUSSION

The Use of the GoFlag Enrichment Set in Orthotricheae

The different analyses performed (Probe Only supermatrix *vs* Full Sequences supermatrix; species-based trees *vs* gene-based trees) yielded overall congruent results and well resolved phylogenetic reconstructions. This adds evidence to the utility of the GoFlag enrichment probe set (Breinholt et al., 2021) to resolve not only phylogenetic relationships across distantly related taxa, but also among more closely related taxa, as it has been demonstrated for other groups of flagellate land plants such as ferns (e.g., Fawcett et al., 2021).

Nevertheless, there was discordance regarding the position of some species depending on the analyses (i.e., topology of the ML supermatrix trees *vs* ASTRAL species trees). There

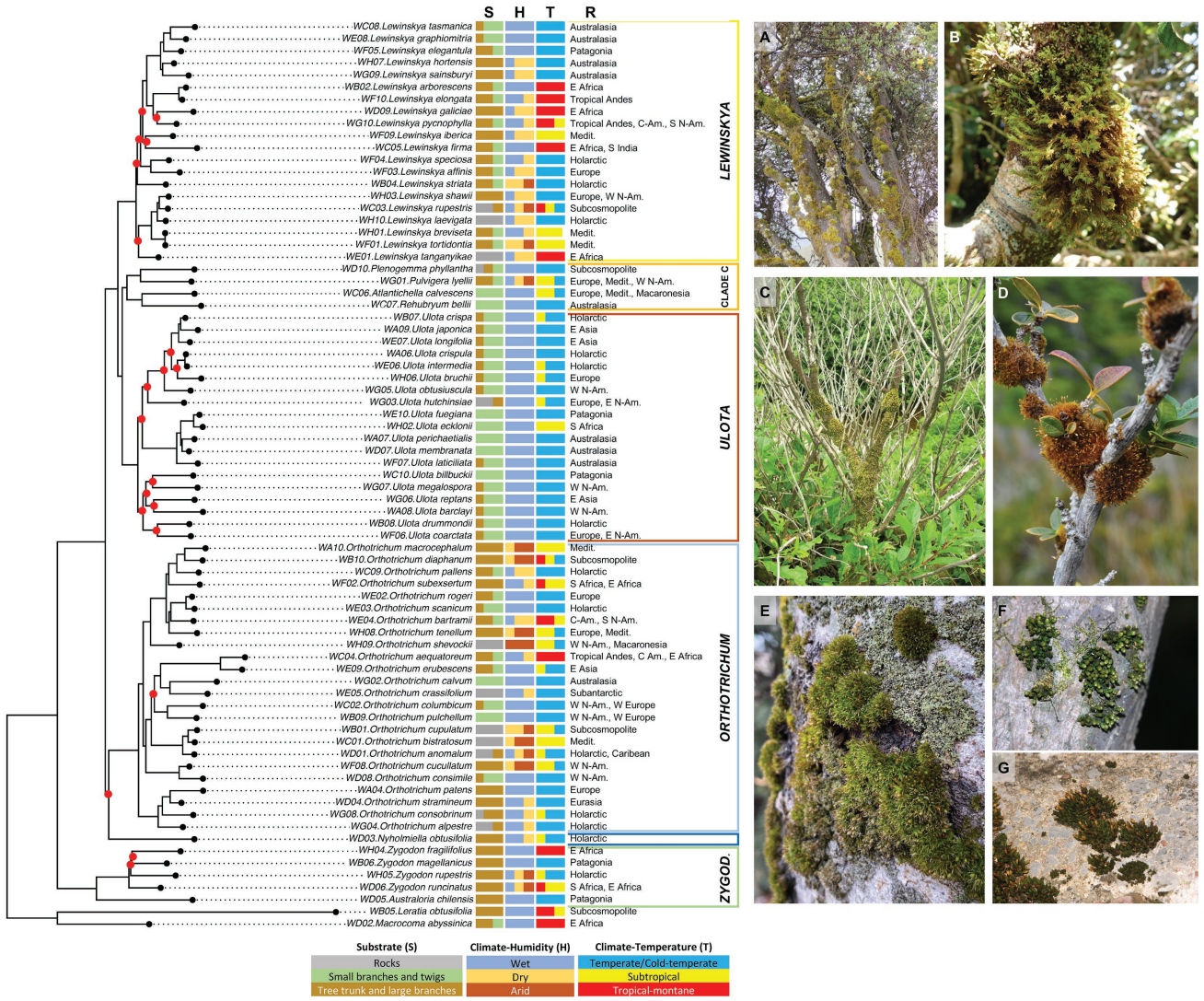

FIGURE 2 | Ecological and biogeographical affinities of the species in the phylogenetic framework (IQ-TREE analysis). For each species, the following information is shown sorted by columns: S—substrate preferences; H—Climate-Humidity, i.e., degree of humidity of the climatic environment in which the species usually grows; T—Climate-Temperature, i.e., degree of climatic thermicity according to the latitudinal climatic bands where the species thrive; R—geographical range. See Materials and Methods section for details on the environmental categories analyzed. Representative pictures: **(A)** *Lewinskya* spp. dominating communities on tree trunks from Bale Mts., Ethiopia; (**B**) *L. graphiomitria* on small branches from Mt. Egmont, North Is., New Zealand; **(C)** *Plenogemma phyllantha* on a shrub trunk and branches from coastal Olympic Peninsula, Washington, United States; and **(D)** *Ulota fuegiana* on shrub twigs from Beagle Channel, Patagonia, Chile. **(E)** *Orthotrichum diaphanum* dominating communities on tree trunk from Madrid, central Spain; **(F)** *O. consobrinum* on a tree trunk from Nara, Honshu Is., Japan; and **(G)** *O. anomalum* on a rock from Burgos, Spain. Abbreviatures: Medit.—Mediterranean Basin, N-Am.—North America; and C-Am.—Central America.

are numerous reasons that can explain this type of conflict, including: (i) short branch lengths (i.e., not enough molecular evidence to be able to resolve the phylogenetic relationships among taxa), (ii) incomplete lineage sorting or another type of genuine conflicting genealogical histories (see a revision in, e.g., Degnan and Rosenberg, 2009), or (iii) a selection of loci that are not informative or that provide unclear information (phylogenetic noise, *sensu* Straub et al., 2014). In our case, most of the nodes that produce conflict are associated with short branches, which is especially evident in the case of the inner nodes of *Lewinskya* (clade B in **Figure 1**) and the position of clade C in relation to clades B and D. The branch suspending the relationship between clade C and clades B-D is very short (due to lack of information, a quick event of diversification, etc.) and thus, some uncertainty involves it: it might appear as sister to either clade B or clade D.

In addition, some of the branches that receive low BS or LPP support have as well low quartet scores, indicating gene tree conflict, and/or have low gCF scores, which reflects that few gene trees support the grouping. For instance, clade E has low BS and LPP support. For this node, which corresponds to a very short branch in both trees, the three quartets have

similar values (q1 = 0.39, q2 = 0.29 and q3 = 0.31) and the gCF/sCF are low (14 and 33%, respectively).

In this scenario, it is necessary to further analyze the poorly supported relationships before making conclusions that involve the conflicting nodes within *Lewinskya*, *Ulota,* and *Orthotrichum*. Future studies will include designing a specific target enrichment set for this family, since this could potentially decrease phylogenetic noise and missing data, as it has been previously demonstrated for other groups such as the tribe Cardueae of Compositae (Herrando-Moraira et al., 2019) or the Cyperaceae (Larridon et al., 2020).

Evolutionary History of the Tribe Orthotricheae

Our phylogenetic reconstruction is the most complete so far published for the tribe Orthotricheae, both in terms of number of taxa included and number of loci analyzed. The results obtained are overall congruent with those published by Draper et al. (2021) and Wang et al. (2021). As proposed by Draper et al. (2021), we confirm that Orthotricheae contains two subtribes, Orthotrichinae F.Lara, Garilleti & Draper and Lewinskyinae, which correspond to clades A (maximally supported) and E (not maximally supported) in **Figure 1**. Moreover, Draper et al. (2021) proposed the segregation at the genus level of both *Australoria* (separate from *Zygodon*, in Zygodonteae) and *Atlantichella* (separate from *Ulota*), which is also supported by our results (**Figure 1**).

According to our phylogenetic reconstructions, the genus *Ulota* as currently conceived remains polyphyletic, since *U. bellii* is placed in a separate fully supported clade (namely C, **Figure 1**) from the rest of the *Ulota* species included in this study (grouped together in a monophyletic and maximally supported clade D in **Figure 1**). *Ulota bellii* shows a characteristic combination of morphological characters that also justifies its segregation from *Ulota* in a separate genus that we propose to name *Rehubryum* F.Lara, Garilleti & Draper. A brief discussion of the diagnostic morphological characters is provided in the taxonomical description section.

None of the phylogenetic reconstructions so far published has been able to fully resolve the relationships of the different genera within the two subtribes, since many of the clades lacked support. In addition, the relationships suggested by previous studies pointed to incongruent results. Based on their 6-loci results, Wang et al. (2021) suggested that, within Lewinskyinae, *Ulota* is sister to *Lewinskya* and that these are grouped with *Plenogemma* and *Pulvigera* in an unresolved polytomy. Noteworthy, these authors did not include the genera *Atlantichella* and *Rehubryum* in their study. Conversely, Draper et al. (2021) considered *Plenogemma* as sister to *Ulota*, and both of them were grouped in a polytomy with *Lewinskya* and *Atlantichella*, based on a selection of 4 different loci and without representation of *Rehubryum*. Regarding Lewinskyinae, we obtained different topologies depending on the analyses, but all our results point to a sister relationship of *Plenogemma* and *Pulvigera*, as well as of *Atlantichella* and *Rehubryum*, and these four genera are assembled in a monophyletic maximally supported clade C (**Figure 1**). The sister relationship of this clade, with either *Ulota* (species-based trees) or *Lewinskya* (gene-based trees), remains ambiguous. Our results fail to provide final evidence regarding the phylogenetic relationships for the genera within Orthotrichinae, since we lack data for *Stoneobryum* D.H.Norris & H.Rob. and *Sehnemobryum* Lewinsky & Hedenäs, so further studies are needed to reach final conclusions. Nevertheless, our results point to a different solution than those suggested in the previously published phylogenies and stress the need to further explore this group of taxa to unravel the intergeneric relationships. On one hand, it is necessary to include all the genera of Orthotricheae in a complete phylogeny to resolve the relationships within Orthotrichinae. On the other hand, there is a need to obtain additional molecular data that could help to discern the evolutionary history of the group. As an example, the conflicting solutions suggested by this study (based on nuclear loci) and those previously published (which include data from organellar genomes) could reflect a complex evolutionary history with ancient hybridization events, as it has been observed in other groups such as algae (e.g., Bringloe et al., 2021), angiosperms (e.g., Bogdanova et al., 2021), or other bryophytes (Meleshko et al., 2021).

In any case, our results indicate a puzzling biogeographic history for the extant taxa, given the strongly supported close relationship of *Plenogemma*, *Pulvigera*, *Atlantichella,* and *Rehubryum* shown by all the analyses (clade C, **Figure 1**). These four taxa include hyperoceanic mosses, but they highly differ in their distributions, reproductive strategies, and morphology (**Figure 2**). *Pulvigera* comprises four species with *Orthotrichum*-like aspect, all of them found in westernmost North America, although one species can also be found in some Pacific archipelagos, and another one is present in western Europe and the Mediterranean (Lara et al., 2020). All species of *Pulvigera* are dioicous mosses with no specialized vegetative reproduction, except for *P. lyellii*, the one with a disjoint Holarctic distribution, which generates abundant gemmae for vegetative propagation. *Plenogemma phyllantha*, the only representative of its genus, is an *Ulota*-like moss with dioicous distribution of sexes that reproduces mainly by vegetative propagules. It shows a wide and irregular bipolar distribution, involving most continents and several oceanic archipelagoes, including some remote sub-Antarctic islands (Garilleti et al., 2015). In turn, both *Atlantichella* and *Rehubryum* are monotypic genera comprising *Ulota*-like, monoicous mosses that actively reproduce sexually and lack any type of specialized propagules for vegetative reproduction. *Atlantichella calvescens* is an endemic of the northeastern Atlantic area, found in the Macaronesian archipelagoes, British Isles, and scattered localities on the western coast of Europe and the Mediterranean basin (Lara et al., 2022), whereas *Rehubryum bellii* is only known from the Antipodes, restricted to New Zealand.

The close relationship of the four taxa comprising clade C agrees with some morphological similarities (see below) but raises a question about the aspect, sexual system, and distribution of their common ancestors and the evolutionary history of the group. The appearance of the ancestral taxon could be either of the two shown by the current descendants, since an original *Ulota*-like appearance would agree with

the topology established by the species tree phylogeny, while the *Orthotrichum*-like appearance would be supported by the topology of the gene-based tree. The two possible sister taxa (*Lewinskya* and *Ulota* s.s.) are entirely monoicous lineages, so the dioicous condition of the subclade formed by *Plenogemma* and *Pulvigera* would in any case be a derived feature, whereas the monoicous condition of the subclade formed by *Atlantichella* and *Rehubryum* would coincide with that of the hypothetic ancestor. We can hypothesize that both the original ancestor and those in the origin of the two main subclades must have been species with high dispersal capacities, as presently shown by many Orthotrichoideae (Vigalondo et al., 2016, 2019). Thanks to recurrent dispersal events, they must have been able to colonize distant hyperoceanic areas of the planet.

In addition to the relationships among the genera of Orthotricheae, this study provides data regarding the infrageneric grouping within the most speciose genera of the tribe. Several infrageneric proposals have been made based on morphological resemblances, all of them focusing exclusively on *Orthotrichum sensu lato* (for a summary, see Lewinsky-Haapasaari and Hedenäs, 1998). Our results suggest that *Lewinskya*, *Ulota,* and *Orthotrichum* can be subdivided into several groups: at least three clades could be recognized within *Lewinskya* (although this grouping is not maximally supported and depends on the analysis performed, species- or gene-based trees); also samples of *Ulota sensu stricto* are distributed in at least three clades, although two of them are not maximally supported; and four groups are established within *Orthotrichum*, all of them maximally supported although their sister relationships partly vary depending on the analyses. Noteworthy, none of these clades is fully congruent with those currently in use (Vitt, 1973; Lewinsky, 1993; Lewinsky-Haapasaari and Hedenas, 1998), although many of the clades here suggested reflect either morphological similarities or ecological preferences (**Figure 2**). As an example, one of these clades unites the most xerophytic taxa included in the analysis (namely, *O. macrocephalum*, *O. diaphanum*, *O. pallens*, *O. subexsertum*, *O. rogeri*, *O. scanicum*, *O. bartramii*, *O. tenellum*, and *O. shevockii*), while another includes taxa that share stomata located in the lower part of the capsule and a hairy vaginula (*O. patens*, *O. stramineum*, *O. alpestre*, and *O. consobrinum*). Similar results pointing that the traditionally accepted subgenera do not reflect natural phylogenetic groups have been obtained in previous studies (e.g., Goffinet et al., 2004; Sawicki et al., 2012). Unfortunately, this study lacks a complete representation of the diversity of *Orthotrichum* (represented here by 26 of 100 species), *Lewinskya* (22/70), and *Ulota* (20/70), and so the present results are too preliminary as to already propose any new infrageneric division. More data are also needed to increase the resolution of the groups and to be able to infer their taxonomic status.

Diversification in the Epiphytic Environment

There is a generally accepted assumption that the epiphytic environment constitutes a hostile one for the development of plants, mainly due to drought stress and restricted nutrient supply (e.g., Pugnaire and Valladares, 1999). Nevertheless, it has been also argued that the epiphytic environment can as well be considered as an available space with unexploited resources (Lüttge, 2008) and with a high diversity of microhabitats due to different gradients of light, temperature, humidity, nutrient supply, and substrate characteristics related to bark structure and branch demography (e.g., Zotz, 2016). This has been especially analyzed in tropical forests (e.g., Lüttge, 2008) and on epiphytic vascular plants (Zotz, 2016). A revision synthesizing the underlying biotic interactions that can have been important for epiphyte ecology and evolution has been recently published (Spicer and Woods, 2022). In this study, the authors highlight the importance of acquiring unique adaptive traits to thrive in fine-scale microhabitats within the epiphytic environment, as evolutive drivers in some vascular epiphyte groups. This has been especially claimed for orchids (e.g., Givnish et al., 2015) and bromeliads (e.g., Benzing, 1987), but little has been published on non-vascular plants (Spicer and Woods, 2022) and the specific factors that promote diversification in mosses are not yet well known.

Bryophytes are poikilohydric organisms whose behavior and adaptations to drought stress strongly differ from those of vascular plants (Barkman, 1958). Mosses and other bryophytes compensate the absence of an impermeabilizing epidermis with the ability to enter in a dormancy state that enables them to tolerate desiccation for long periods, whereas water uptake is mostly ectohydric. According to Huttunen et al. (2018), morphological characteristics connected to ectohydry may be driven by adaptations to environmental conditions. This could be interpreted as evidence of how acquiring adaptive traits can drive diversification in bryophytes, which has been suggested for orchids and bromeliads.

Orthotrichaceae is one of the most speciose bryophyte families and has diversified mostly in the epiphytic environment. Within it, the very species-rich subfamily Macromitrioideae has diversified in warm tropical epiphytic environments where also most vascular epiphytes grow. But to what extent the diversification of the tribe Orthotricheae, which specialized in temperate environments (including high tropical altitudes), can be interpreted in the same terms of adaptation to a wide variety of meso- and microenvironmental conditions is something that has not been previously addressed. Previous works lacked support to approach this question, but our results have revealed a clear ecological pattern involving substrate and climatic preferences in the major clades recovered that make progress toward identifying possible drivers of diversification. Conversely, we have not been able to detect a clear biogeographical signal in the phylogenetic reconstructions. This could indicate the evolutionary importance of acquiring adaptative traits that enable the colonization of certain epiphytic microhabitats. In this context, the diversification at the genus level in the tribe Orthotricheae could be partially explained by the adaptation to a certain combination of the substrate characteristics (rocks, trunks, or small branches) together with climatic preferences regarding humidity and temperature. This could also explain an infrageneric diversification, although further data and the inclusion of a wider representation of species are needed to be able to safely achieve conclusions at this taxonomical level.

Finally, as suggested by Huttunen et al. (2018) and in line with the results by Draper et al. (2021) on the prevalence of homoplasy among Orthotricheae, the importance of the adaptation to the environment could explain the parallel morphological evolution of the different genera specialized on the epiphytic habitat. Moreover, the results of our study point toward the idea that, at least for bryophytes, stressful environments can promote diversification and harbor great diversity.

Taxonomical Description

The genus *Rehubryum* is proposed to accommodate *Ulota bellii* Malta on the basis of its peculiar combination of morphological traits and phylogenetic position.

Rehubryum F.Lara, Garilleti & Draper, *gen. nov.*

Type: *Rehubryum bellii* (Malta) F.Lara, Garilleti & Draper, comb. nov. ≡ *Ulota bellii* Malta, Acta Horti Bot. Univ. Latv. 7: 15. 1933.

Diagnosis: Plants autoicous, forming cushions. Leaves spirally arranged, strongly crisped when dry, lanceolate, gradually dilated to a base scarcely concave, often plicate on both sides of the nerve, with margins plane or erect-incurved in one side, especially in the transition between base and lamina, leaf lamina unistratose and mainly plane at margins; basal cells long rectangular to linear, somewhat sinuous, with thickened walls; basal-marginal cells differentiated, hyaline, quadrate to rectangular, with thickened transverse walls, forming a narrow marginal band along the base and proximal end of the lamina; margins at upper base with papillose teeth arising at the junctions between two cells; submarginal rows of elongated cells differentiated from base through lower third of the lamina; median and upper leaf-cells rounded to elliptical, with low papillae. Propagula absent. Perichaetial leaves slightly differentiated, with a broader base. Seta 3–5 mm long, twisted counterclockwise. Capsule exserted, oblong-ovoid to short cylindrical, symmetric, entirely ribbed. Exothecial bands narrow differentiated from mouth to urn base. Stomata superficial, at urn base and neck. Peristome double; exostome of 8 pairs of teeth, easily splitting after recurving; endostome of 16 linear segments, involute when dry, with a low connective membrane. Operculum rostrate, with base almost plane. Spores unicellular, isomorphic, papillose. Calyptra mitrate, with abundant stout hairs.

Etymology: *rehu* is a Maori noun for mist but also a verb that means to pass out of sight, disappear, and render unconscious (Moorfield, 2022), all of which seems appropriate for this moss that lives in foggy environments and has gone virtually unnoticed as a different genus.

The New Zealand endemic *Rehubryum bellii* appears to be a typical species of the genus *Ulota* as it shows the general look that most of these mosses have (**Figures 3A,B**), as well as many of the details that serve as morphological characters for their taxonomic characterization. Indeed, in his recent review of *Ulota* in New Zealand, Fife (2017) considers *U. bellii* not worthy of taxonomic recognition and synonymized it with *U. lutea* (Hook. f. & Wilson) Mitt. However, *R. bellii* is easily separated from any species of *Ulota* in the Australasian area by the possession of an endostome consisting of 16 filiform segments, involute when dry, all of them well developed (**Figure 3C**). Other differential characters, such as the oblong-ovoid shape of the capsule (**Figure 3B**) or the possession of an exostome with 8 pairs of teeth easily splitting (**Figure 3C**), have already been highlighted since the description of the species (Malta, 1933; Sainsbury, 1955). However, this moss has two additional very significant characters at the distal part of the leaf base (**Figure 3D**): (a) submarginal bands of elongate cells ascending from the transition base-blade some way up and (b) margins of some leaves denticulated by prominent papillae arising at the junction between every two marginal cells. Both characters seem to have gone unnoticed and their discovery while examining our New Zealand collections was fundamental for the inclusion of samples of this species in the phylogenetic analysis. In fact, in a previous study (Draper et al., 2021), the occurrence of leaves with submarginal bands of elongate cells was revealed as a characteristic shared by the genera *Atlantichella* and *Plenogemma*, whereas basal leaf margins with geminate teeth are also found in these two genera and in *Pulvigera*, where the feature is especially visible. In contrast, both traits appear to be absent in *Ulota*. The phylogenetic reconstructions obtained in the present study group in the same clade all the four genera in which these features have been so far observed, giving these characteristics an unsuspected taxonomic significance.

The distinction between *Rehubryum bellii* and *Atlantichella calvescens* does not entail any difficulty because there are many

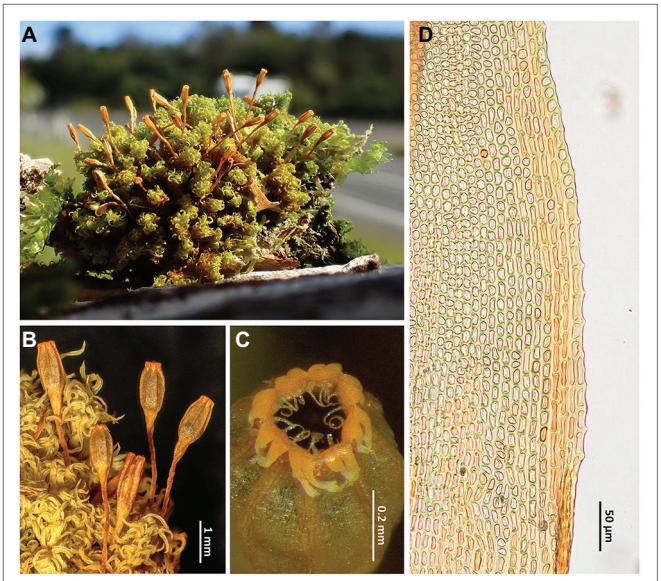

FIGURE 3 | *Rehubryum bellii*. **(A)** General aspect of a dry cushion in the field. **(B)** Detail of the habit showing the upper leaves when dry and several sporophytes, most of them with mature, recently opened capsules. **(C)** Mouth of a capsule when dry, showing a peristome with 8 pairs of teeth, split almost to the base, and 16 well-developed segments. **(D)** Detail of the leaf lamina margin just above the basal leaf showing the submarginal band of elongated cells and the papillose marginal cells. **(A)** image taken in Mount Taranaki NP (Lara 1601/57, MAUAM); **(A–C)** from Garilleti 2016-045b (Garilleti's personal Herb.); and D from Garilleti 2016–104 (Garilleti's personal Herb.).

morphological differences between the two species, especially in the sporophyte. Thus, for example, whereas *A. calvescens* has capsules broadly ribbed and strongly contracted below mouth when dry, with an endostome of 8 linear segments and devoid of connective membrane, *R. bellii* has capsules finely ribbed, not contracted below mouth when dry, with an endostome of 16 filiform segments and with connective membrane. As these are the only known species of these two genera, it could be thought that their differential traits also serve to characterize *Rehubryum* versus *Atlantichella*. However, all the morphological characters that serve for differentiating both species vary within the large genus *Ulota* (Caparrós et al., 2014; Caparrós, 2015), so their value for characterizing genera among the Lewinskyinae is doubtful. It should also be noted that Draper et al. (2021) demonstrated that within this group of mosses most characters used for separating genera are homoplastic.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repositories and accession numbers or links can be found at Figshare: https://figshare.com/s/bb78149766ac8809b4fc; NCBI: PRJNA819401.

AUTHOR CONTRIBUTIONS

ID, FL, and RG designed the research. ID, FL, RG, VM, and JC sampled and selected the specimens for the molecular analyses. FL and RG selected and processed the specimens for the morphological study. GB and SM processed the specimens and obtained the sequences. TV, ID, FL, GB, and SM contributed to the phylogenetical analyses. RG, TV, ID, and FL prepared the illustrations. FL, RG, and ID performed the analyses regarding the ecological and geographical preferences. ID and FL wrote a first draft of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by the Spanish Ministry of Economy, Industry and Competitiveness (grant CGL2016-80772-P), the Spanish Research Agency of the Ministry of Science and Innovation (PID2020-115149GB-C21 and PID2020-115149GB-C22), and the U.S. National Science Foundation (DEB-1541506).

ACKNOWLEDGMENTS

The authors deeply thank the editor G. M. Schneeweiss and two reviewers for their valuable suggestions on a previous 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.2022.882960/full#supplementary-material

REFERENCES

Aigoin, D. A., Devos, N., Huttunen, S., Ignatov, M. S., Gonzalez-Mancebo, J. M., and Vanderpoorten, A. (2009). And if Engler was not completely wrong? Evidence for multiple evolutionary origins in the moss flora of Macaronesia. *Evolution* 63, 3248–3257. doi: 10.1111/j.1558-5646.2009.00787.x

Barkman, J. J. (1958). *Phytosociology and Ecology of Cryptogamic Epiphytes. Including a Taxonomic Survey and Description of their Vegetation Units in Europe*. Assen: Van Gorcum.

Benzing, D. H. (1987). Vascular epiphytism: taxonomic participation and adaptive diversity. *Ann. Mo. Bot. Gard.* 74, 183–204. doi: 10.2307/2399394

Bogdanova, V. S., Shatskaya, N. V., Mglinets, A. V., Kosterin, O. E., and Vasiliev, G. V. (2021). Discordant evolution of organellar genomes in peas (Pisum L.). *Mol. Phylogenet. Evol.* 160:107136. doi: 10.1016/j.ympev.2021.107136

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

Breinholt, J. W., Carey, S. B., Tiley, G. P., Davis, E. C., Endara, L., McDaniel, S. F., et al. (2021). A target enrichment probe set for resolving the flagellate land plant tree of life. *Appl. Plant Sci.* 9:e11406. doi: 10.1002/aps3.11406

Breinholt, J. W., Carey, S. B., Tiley, G. P., Davis, E. C., Endara, L., McDaniel, S. F., et al. (2020). A target enrichment probe set for resolving the flagellate land plant tree of life. 49511638 bytes. doi: 10.5061/DRYAD.7PVMCVDQG.

Breinholt, J. W., Earl, C., Lemmon, A. R., Lemmon, E. M., Xiao, L., and Kawahara, A. Y. (2018). Resolving relationships among the megadiverse butterflies and moths with a novel pipeline for anchored phylogenomics. *Syst. Biol.* 67, 78–93. doi: 10.1093/sysbio/syx048

Bringloe, T. T., Zaparenkov, D., Starko, S., Grant, W. S., Vieira, C., Kawai, H., et al. (2021). Whole-genome sequencing reveals forgotten lineages and recurrent hybridizations within the kelp genus *Alaria* (Phaeophyceae). *J. Phycol.* 57, 1721–1738. doi: 10.1111/jpy.13212

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

Caparrós, R. (2015). El género *Ulota* D. Mohr en la Península Ibérica y una nueva visión del complejo de *U. crispa* (Hedw.) Brid. (Orthotrichaceae, Musci). Available at: https://www.educacion.gob.es/teseo (Accessed April 1, 2022).

Caparrós, R., Garilleti, R., and Lara, F. (2014). "*Ulota* D.Mohr," in *Flora Briofítica Ibérica, vol. V. Orthotrichales: Orthotrichaceae; Hedwigiales: Hedwigiaceae; Leucodontales: Fontinalaceae, Climaciaceae, Anomodontaceae, Cryphaeaceae, Leptodontaceae, Leucodontaceae, Neckeraceae; Hookeriales: Hypopterygiaceae, Hookeriaceae, Leucomiaceae, Pilotrichaceae*. eds. J. Guerra, M. J. Cano and M. Brugués (Murcia: Universidad de Murcia & Sociedad Española de Briología), 34–50.

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

Chang, Z., Li, G., Liu, J., Zhang, Y., Ashby, C., Liu, D., et al. (2015). Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. *Genome Biol.* 16:30. doi: 10.1186/s13059-015-0596-2

Degnan, J. H., and Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* 24, 332–340. doi: 10.1016/j.tree.2009.01.009

Dong, S., and Liu, Y. (2021). The mitochondrial genomes of bryophytes. *Bryophyte Divers. Evol.* 43, 112–126. doi: 10.11646/bde.43.1.9

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

Draper, I., Garilleti, R., Calleja, J. A., Flagmeier, M., Mazimpaka, V., Vigalondo, B., et al. (2021). Insights into the evolutionary history of the subfamily Orthotrichoideae (Orthotrichaceae, Bryophyta): new and former supra-specific taxa so far obscured by prevailing homoplasy. *Front. Plant Sci.* 12:629035. doi: 10.3389/fpls.2021.629035

Draper, I., Lara, F., Albertos, B., Garilleti, R., and Mazimpaka, V. (2006). Epiphytic bryoflora of the Atlas and Antiatlas Mountains, including a synthesis of the distribution of epiphytic bryophytes in Morocco. *J. Bryol.* 28, 312–330. doi: 10.1179/174328206x136313

Fawcett, S., Smith, A. R., Sundue, M., Burleigh, J. G., Sessa, E. B., Kuo, L.-Y., et al. (2021). A global phylogenomic study of the Thelypteridaceae. *Syst. Bot.* 46, 891–915. doi: 10.1600/036364421X16370109698650

Fife, A. J. (2017). "Orthotrichaceae," in *Flora of New Zealand. Mosses. Fascicle 31*. eds. I. Breitwieser and A. D. Wilton (Lincoln: Manaaki Whenua Press), 113.

Frey, W., and Stech, M. (2009). "Division of Bryophyta Schimp. (Musci, Mosses)," in *Syllabus of Plant Families. Adolf Engler's Syllabus der Pflanzenfamilien, 13th edition. Part 3. Bryophytes and Seedless Vascular Plants*. ed. W. Frey (Berlin: Gebrüder Borntraeger), 116–257.

Garilleti, R., Mazimpaka, V., and Lara, F. (2015). *Ulota larrainii* (Orthotrichoideae, Orthotrichaceae, Bryophyta) a new species from Chile, with comments on the worldwide diversification of the genus. *Phytotaxa* 217, 133–144. doi: 10.11646/phytotaxa.217.2.3

Givnish, T. J., Spalink, D., Ames, M., Lyon, S. P., Hunter, S. J., Zuluaga, A., et al. (2015). Orchid phylogenomics and multiple drivers of their extraordinary diversification. *Proc. R. Soc. B Biol. Sci.* 282:20151553. doi: 10.1098/rspb.2015.1553

Goffinet, B., Shaw, A. J., Cox, C. J., Wickett, N. J., and Boles, S. B. (2004). Phylogenetic inferences in the Orthotrichoideae (Orthotrichaceae, Bryophyta) based on variation in four loci from all genomes. *Monogr. Syst. Bot. Mo. Bot. Gard.* 98, 270–289.

Goffinet, B., and Vitt, D. H. (1998). "Revised generic classification of the Orthotrichaceae based on a molecular phylogeny and comparative morphology," in *Bryology for the Twenty-First Century*. eds. J. W. Bates, N. W. Ashton and J. G. Duckett (Leeds: Maney Publishing and the British Bryological Society), 143–160.

Herrando-Moraira, S., Calleja, J. A., Galbany-Casals, M., Garcia-Jacas, N., Liu, J.-Q., López-Alvarado, J., et al. (2019). Nuclear and plastid DNA phylogeny of tribe Cardueae (Compositae) with Hyb-Seq data: A new subtribal classification and a temporal diversification framework. *Mol. Phylogenet. Evol.* 137, 313–332. doi: 10.1016/j.ympev.2019.05.001

Huttunen, S., Bell, N., and Hedenäs, L. (2018). The evolutionary diversity of mosses – taxonomic heterogeneity and its ecological drivers. *Crit. Rev. Plant Sci.* 37, 128–174. doi: 10.1080/07352689.2018.1482434

IPNI (2021). International Plant Names Index. The Royal Botanic Gardens, Kew, Harvard University Herbaria & Libraries and Australian National Botanic Gardens. Available at: http://www.ipni.org. (Accessed February 1, 2022).

Junier, T., and Zdobnov, E. M. (2010). The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinforma. Oxf. Engl.* 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 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

Lara, F., Garilleti, R., Mazimpaka, V., and Guerra, J. eds. (2014). "Orthotrichaceae," in *Flora Briofítica Ibérica, vol. V. Orthotrichales: Orthotrichaceae; Hedwigiales: Hedwigiaceae; Leucodontales: Fontinalaceae, Climaciaceae, Anomodontaceae, Cryphaeaceae, Leptodontaceae, Leucodontaceae, Neckeraceae; Hookeriales: Hypopterygiaceae, Hookeriaceae, Leucomiaceae, Pilotrichaceae* (Murcia: Universidad de Murcia & Sociedad Española de Briología), 15–135.

Lara, F., Draper, I., Flagmeier, M., Calleja, J. A., Mazimpaka, V., and Garilleti, R. (2020). Let's make *Pulvigera* great again: re-circumscription of a misunderstood group of Orthotrichaceae that diversified in North America. *Bot. J. Linn. Soc.* 193, 180–206. doi: 10.1093/botlinnean/boaa013

Lara, F., Garilleti, R., Goffinet, B., Draper, I., Medina, R., Vigalondo, B., et al. (2016). *Lewinskya*, a new genus to accommodate the phaneroporous and monoicous taxa of *Orthotrichum* (Bryophyta, Orthotrichaceae). *Cryptogam. Bryol.* 37, 361–382. doi: 10.7872/cryb/v37.iss4.2016.361

Lara, F., Garilleti, R., Medina, R., and Mazimpaka, V. (2009). A new key to the genus *Orthotrichum* in Europe and the Mediterranean region. *Cryptogam. Bryol.* 30, 129–142.

Lara, F., Porley, R. D., Draper, I., Aleffi, M., Garcia, C., and Garilleti, R. (2022). *Ulota* s.l. (Orthotrichaceae, Bryidae) at shouthernmost Mediterranean localities: not a simple matter. *Plant Biosyst.* 1–8. doi: 10.1080/11263504.2022.2056652 [Epub ahead of print].

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., and Lemmon, A. (2013). High-throughput genomic data in systematics and phylogenetics. *Annu. Rev. Ecol. Evol. Syst.* 44, 99–121. doi: 10.1146/annurev-ecolsys-110512-135822

Lewinsky, J. (1993). A synopsis of the genus *Orthotrichum* Hedw. (Musci: Orthotrichaceae). *Bryobrothera* 2, 1–59.

Lewinsky-Haapasaari, J., and Hedenäs, L. (1998). A cladistic analysis of the genus *Orthotrichum*. *Bryologist* 101, 519–555. doi: 10.1639/0007-2745(1998)101[519:ACAOTM]2.0.CO;2

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, 1–11. doi: 10.1038/s41467-019-09454-w

Liu, Y., Medina, R., and Goffinet, B. (2014). 350 My of mitochondrial genome stasis in mosses, an early land plant lineage. *Mol. Biol. Evol.* 31, 2586–2591. doi: 10.1093/molbev/msu199

Lüttge, U. (2008). *Physiological Ecology of Tropical Plants. 2nd Edn.* Berlin: Springer.

Malta, N. (1933). A survey of the Australasian species of *Ulota*. *Acta Horti Bot. Univ. Latv.* 7, 1–24.

Mazimpaka, V., and Lara, F. (1995). Corticolous bryophytes of *Quercus pyrenaica* forests from Gredos Mountains (Spain): vertical distribution and affinity for epiphytic habitats. *Nova Hedwig.* 61, 431–446.

McDaniel, S. F., and Shaw, A. J. (2003). Phylogeographic structure and cryptic speciation in the trans-Antarctic moss *Pyrrhobryum mnioides*. *Evolution* 57, 205–215. doi: 10.1554/0014-3820(2003)057[0205:PSACSI]2.0.CO;2

Meleshko, O., Martin, M. D., Korneliussen, T. S., Schröck, C., Lamkowski, P., Schmutz, J., et al. (2021). Extensive genome-wide phylogenetic discordance is due to incomplete lineage sorting and not ongoing introgression in a rapidly radiated bryophyte genus. *Mol. Biol. Evol.* 38, 2750–2766. doi: 10.1093/molbev/msab063

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., et al. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37, 1530–1534. doi: 10.1093/molbev/msaa015

Moorfield, J. C. (2022). *Te Aka Māori Dictionary*. Available at: https://maoridictionary.co.nz (Accessed February 20, 2022).

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

Plášek, V., Sawicki, J., Ochyra, R., Szczecińska, M., and Kulik, T. (2015). New taxonomical arrangement of the traditionally conceived genera *Orthotrichum* and *Ulota* (Orthotrichaceae, Bryophyta). *Acta Musei Silesiae Sci. Nat.* 64, 169–174. doi: 10.1515/cszma-2015-0024

Pugnaire, F. I., and Valladares, F. (1999). *Handbook of Functional Plant Ecology*. New York: Marcel Dekker, Inc.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R foundation for Statistical Computing.

Renner, M. A. M. (2020). Opportunities and challenges presented by cryptic bryophyte species. *Telopea* 23, 41–60. doi: 10.7751/telopea14083

Rosato, M., Kovařík, A., Garilleti, R., and Rosselló, J. A. (2016). Conserved organisation of 45S rDNA sites and rDNA gene copy number among major clades of early land plants. *PLoS One* 11:e0162544. doi: 10.1371/journal.pone.0162544

Sainsbury, G. O. K. (1955). A handbook of New Zealand mosses. *Bull. R. Soc. N. Z.* 5, 1–490.

Sawicki, J., Plášek, V., and Szczecińska, M. (2012). Molecular data do not support the current division of *Orthotrichum* (Bryophyta) species with immersed stomata. *J. Syst. Evol.* 50, 12–24. doi: 10.1111/j.1759-6831.2011.00168.x

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

Shah, T., Schneider, J. V., Zizka, G., Maurin, O., Baker, W., Forest, F., et al. (2021). Joining forces in Ochnaceae phylogenomics: a tale of two targeted sequencing probe kits. *Am. J. Bot.* 108, 1201–1216. doi: 10.1002/ajb2.1682

Spicer, M. E., and Woods, C. L. (2022). A case for studying biotic interactions in epiphyte ecology and evolution. *Perspect. Plant Ecol. Evol. Syst.* 54:125658. doi: 10.1016/j.ppees.2021.125658

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

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

Troll, C., and Paffen, K. H. (1964). Karte der Jahreszeiten-Klimate der Erde. *Erdkunde* 18, 5–28. doi: 10.3112/erdkunde.1964.01.02

Tropicos (2022) Home. Available at: https://www.tropicos.org/home (Accessed February 17, 2022).

Vigalondo, B., Lara, F., Draper, I., Valcárcel, V., Garilleti, R., and Mazimpaka, V. (2016). Is it really you, *Orthotrichum acuminatum*? Ascertaining a new case of intercontinental disjunction in mosses. *Bot. J. Linn. Soc.* 180, 30–49. doi: 10.1111/boj.12360

Vigalondo, B., Patiño, J., Draper, I., Mazimpaka, V., Shevock, J. R., Losada-lima, A., et al. (2019). The long journey of *Orthotrichum shevockii* (Orthotrichaceae, Bryopsida): from California to Macaronesia. *PLoS One* 14:e0211017. doi: 10.1371/journal.pone.0211017

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

Vitt, D. H. (1973). A revision of the genus *Orthotrichum* in North America, north of Mexico. *Bryophyt. Bibl.* 1, 1–165.

Wang, Q.-H., Dong, S.-S., Zhang, J.-L., Liu, Y., and Jia, Y. (2021). Phylogeny of *Orthotrichum* s.l. and *Ulota* s.l. (Orthotrichaceae, Bryophyta): insights into stomatal evolution. *J. Syst. Evol.* doi: 10.1111/jse.12713

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

Yu, G. (2021). *ggimage: Use Image in "ggplot2."* Available at: https://CRAN.R-project.org/package=ggimage (Accessed February 1, 2022).

Yu, G., Smith, D. K., Zhu, H., Guan, Y., and Lam, T. T.-Y. (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

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

Zotz, G. (2016). *Plants on Plants: The Biology of Vascular Epiphytes*. Switzerland: Springer.

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

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

Copyright © 2022 Draper, Villaverde, Garilleti, Burleigh, McDaniel, Mazimpaka, Calleja and Lara. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 20 May 2022
doi: 10.3389/fpls.2022.870949

Deep Insights Into the Plastome Evolution and Phylogenetic Relationships of the Tribe Urticeae (Family Urticaceae)

Catherine A. Ogoma[1,2], Jie Liu[1,3], Gregory W. Stull[1], Moses C. Wambulwa[1,3,4], Oyetola Oyebanji[1,2], Richard I. Milne[5], Alexandre K. Monro[6], Ying Zhao[1], De-Zhu Li[1] and Zeng-Yuan Wu[1]**

[1] Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, China, [2] University of Chinese Academy of Sciences, Beijing, China, [3] Key Laboratory of Biodiversity and Biogeography, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, China, [4] Department of Life Sciences, School of Science and Computing, South Eastern Kenya University, Kitui, Kenya, [5] School of Biological Sciences, Institute of Molecular Plant Sciences, University of Edinburgh, Edinburgh, United Kingdom, [6] Royal Botanic Gardens, Kew, Richmond, United Kingdom

OPEN ACCESS

Edited by:
Ruslan Kalendar,
University of Helsinki, Finland

Reviewed by:
Xue-jun Ge,
South China Botanical Garden (CAS), China
Mark Fishbein,
Oklahoma State University, United States

***Correspondence:**
De-Zhu Li
DZL@mail.kib.ac.cn
Zeng-Yuan Wu
wuzengyuan@mail.kib.ac.cn

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

Received: 07 February 2022
Accepted: 15 April 2022
Published: 20 May 2022

Citation:
Ogoma CA, Liu J, Stull GW, Wambulwa MC, Oyebanji O, Milne RI, Monro AK, Zhao Y, Li D-Z and Wu Z-Y (2022) Deep Insights Into the Plastome Evolution and Phylogenetic Relationships of the Tribe Urticeae (Family Urticaceae). Front. Plant Sci. 13:870949. doi: 10.3389/fpls.2022.870949

Urticeae *s.l.*, a tribe of Urticaceae well-known for their stinging trichomes, consists of more than 10 genera and approximately 220 species. Relationships within this tribe remain poorly known due to the limited molecular and taxonomic sampling in previous studies, and chloroplast genome (CP genome/plastome) evolution is still largely unaddressed. To address these concerns, we used genome skimming data—CP genome and nuclear ribosomal DNA (18S-ITS1-5.8S-ITS2-26S); 106 accessions—for the very first time to attempt resolving the recalcitrant relationships and to explore chloroplast structural evolution across the group. Furthermore, we assembled a taxon rich two-locus dataset of *trnL-F* spacer and ITS sequences across 291 accessions to complement our genome skimming dataset. We found that Urticeae plastomes exhibit the tetrad structure typical of angiosperms, with sizes ranging from 145 to 161 kb and encoding a set of 110–112 unique genes. The studied plastomes have also undergone several structural variations, including inverted repeat (IR) expansions and contractions, inversion of the *trnN-GUU* gene, losses of the *rps19* gene, and the *rpl2* intron, and the proliferation of multiple repeat types; 11 hypervariable regions were also identified. Our phylogenomic analyses largely resolved major relationships across tribe Urticeae, supporting the monophyly of the tribe and most of its genera except for *Laportea*, *Urera*, and *Urtica*, which were recovered as polyphyletic with strong support. Our analyses also resolved with strong support several previously contentious branches: (1) *Girardinia* as a sister to the *Dendrocnide-Discocnide-Laportea-Nanocnide-Zhengyia-Urtica-Hesperocnide* clade and (2) *Poikilospermum* as sister to the recently transcribed *Urera sensu stricto*. Analyses of the taxon-rich, two-locus dataset showed lower support but was largely congruent with results from the CP genome and nuclear ribosomal DNA dataset. Collectively, our study highlights the power of genome skimming data to ameliorate phylogenetic resolution and provides new insights into phylogenetic relationships and chloroplast structural evolution in Urticeae.

Keywords: Urticaceae *s.l.*, chloroplast structural evolution, phylogenomic, genome skimming, Urticaceae

INTRODUCTION

Urticaceae, commonly known as the nettle family, is a cosmopolitan group of plants comprising approximately 54 genera and ∼2,600 species circumscribed into six tribes (Boehmerieae Gaudich., Cecropiaceae Gaudich., Elatostemateae Gaudich., Forsskaoleae Gaudich., Parietarieae Gaudich., and Urticeae Lam. and DC.; Conn and Hadiah, 2009) with various distinct morphological characters (Stevens, 2017). For example, members of tribe Urticeae are well known for their stinging trichomes (Friis, 1993). Urticeae *sensu* Friis (1989) consists of 10 genera of vast economic importance as sources of fiber (Singh and Shrestha, 1988; Bodros and Baley, 2008; Gurung et al., 2012) medicine (Momo et al., 2006; Tanti et al., 2010; Luo et al., 2011; Benvenutti et al., 2020; Sharan Shrestha et al., 2020), and food (Di Virgilio et al., 2015; Mahlangeni et al., 2020). This generic circumscription of the Urticeae, however, was established prior to the era of molecular phylogenetics. With the advent of the molecular tools, classification within tribe Urticeae has received much attention, with both taxonomic and phylogenetic studies spurring realignments (Hadiah et al., 2008; Kim et al., 2015; Huang et al., 2019; Wells et al., 2021). Molecular analyses have led to the exclusion of *Gyrotaenia* and the inclusion of *Touchardia*, *Poikilospermum* and *Zhengyia* in the tribe; hence, Urticeae presently comprises 12 genera (Wu et al., 2013; Kim et al., 2015; Jin et al., 2019). Molecular phylogenetic studies have also been able to demonstrate the monophyly of this tribe as well as which genera are polyphyletic or monophyletic.

Although our understanding of evolutionary relationships of the tribe Urticeae has improved in recent years, some important nodes remain unresolved. For example, the phylogenetic position of *Laportea* remains contentious in previous studies. Wu et al. (2013), using seven combined markers from the mitochondrial, nuclear, and chloroplast genomes, recovered *Laportea* sister to a clade comprising *Obetia-Urera-Touchardia* and *Poikilospermum*, though with weak support (**Figure 1A**). Subsequent studies, however, have supported alternative, conflicting resolutions of *Laportea* (**Figures 1B–D**; Kim et al., 2015; Wu et al., 2018; Huang et al., 2019) probably due to the limited sampling. The placement of *Poikilospermum* also remains uncertain; although it has consistently been placed sister to *Urera*, support for this was either lacking (**Figures 1A–C**; Wu et al., 2013, 2018; Kim et al., 2015; Wells et al., 2021) or low (**Figure 1D**; Huang et al., 2019). The genus *Hesperocnide*, although supported as monophyletic in earlier studies, was recently recovered as polyphyletic by Huang et al. (2019), suggesting that further investigation of this genus may be required. Conflict concerning the placement of *Girardinia* further compounds taxonomic problems within Urticeae; several studies support its relationship with *Dendrocnide-Discocnide*, but without support (**Figures 1A,B**; Wu et al., 2013; Kim et al., 2015), while others (Wu et al., 2018; Huang et al., 2019) have recovered *Girardinia* sister to a clade comprising *Dendrocnide-Discocnide-Laportea-Nanocnide-Zhengyia-Urtica-Hesperocnide,* albeit also with low support (**Figures 1C,D**). These uncertainties around phylogenetic relationships within Urticeae are likely due to limited taxon or genic sampling in previous studies. Therefore, a broadly sampled phylogenomic study should offer useful framework for resolving these outstanding problems and guiding revised taxonomic treatments of the tribe.

Chloroplasts are ubiquitous organelles in plants with tractable attributes that make them highly suitable for use in phylogenetic and phylogeographic studies (Demenou et al., 2020; Silverio et al., 2021; Simmonds et al., 2021; Wang et al., 2021). In Urticaceae, whole chloroplast genomes have proven to be indispensable for sequence variation exploration (Wang et al., 2020b; Li et al., 2021). More broadly, studies of chloroplast genomes have been useful for understanding molecular evolutionary patterns of gene duplication, loss, rearrangement, and transfer across angiosperms (Yan et al., 2018; Do et al., 2020; Liu et al., 2020a; Oyebanji et al., 2020), though discordant relationships may be caused by plastid capture and other evolutionary processes.

For the present study, we sequenced and examined chloroplast genomes (CP genome/plastome) of the tribe Urticeae in order to explore plastome structural evolution in the tribe and to reconstruct the first-ever full plastome phylogeny for the tribe. Furthermore, we generated a robustly sampled dataset of Urticeae (comprising 291 accessions) aimed at reconstructing a more taxonomically rich phylogeny for the tribe. Specifically, we aimed to (1) characterize structural changes in Urticeae plastomes, (2) resolve deep relationships in the tribe using different data partitioning strategies, and (3) evaluate and update existing classifications for Urticeae in the light of our phylogenetic results based on both plastome and nuclear data.

MATERIALS AND METHODS

Taxon Sampling

In this study, we sampled a total of 106 accessions, comprising 90 ingroup accessions (58 spp. in 12 genera) from the tribe Urticeae, plus 12 accessions (12 spp. in 11 genera) from other Urticaceae tribes and four (3 spp. in 3 genera) from outside the family as outgroups. These represent the genome skimming—CP genome and the nuclear ribosomal DNA (18S-ITS1-5.8S-ITS2-26S) dataset for the phylogenetic analyses (**Supplementary Table 1**). Of the 106 accessions, 57 representative accessions (each a different taxon) were selected for CP genome structural analyses. To produce a more comprehensive phylogenetic framework for the tribe Urticeae, we also generated a new two-locus dataset of 291 accessions (145 spp. in 26 genera) based on ITS and the *trnL-F* intergenic spacer. The ITS and the *trnL-F* intergenic spacer dataset was sampled based on maximum taxon data availability on NCBI database. Of the 291 accessions included, 187 sequences were obtained from NCBI GenBank while the remaining were newly sequenced for this study. Information on the plant material (collection localities and voucher specimen numbers) and the associated GenBank accessions are listed in **Supplementary Table 1**.

DNA Extraction and Sequencing

A modified cetyl trimethyl ammonium bromide (CTAB) protocol (Doyle and Doyle, 1987) was used to extract total DNA from both

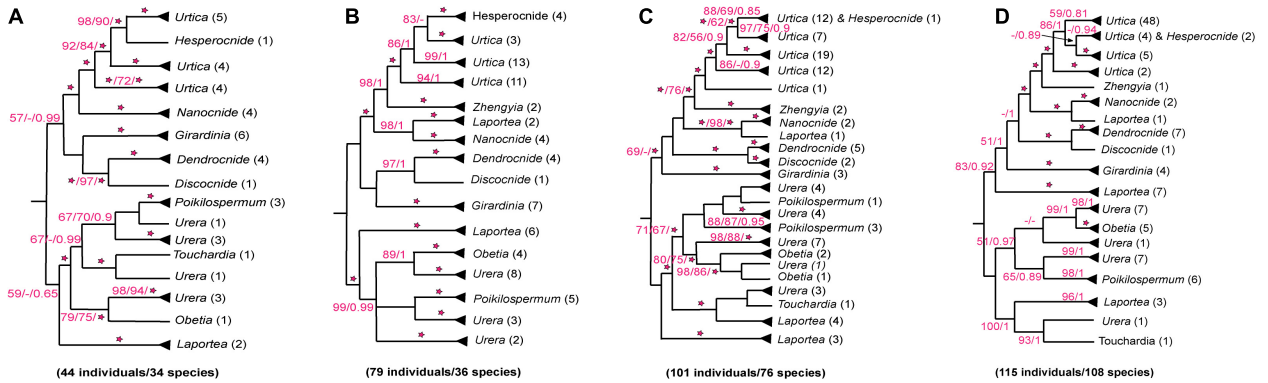

FIGURE 1 | Alternative relationships of Urticeae based on combined loci data from previous *analyses*. **(A)** Wu et al. (2013): Maximum likelihood (ML)/Maximum parsimony (MP)/Bayesian inference (BI) analyses based on nuclear, chloroplast, and mitochondrial loci; **(B)** Kim et al. (2015): MP/BI tree inferred from chloroplast and nuclear DNA data; **(C)** Wu et al. (2018): ML/MP/BI tree inferred from nuclear, chloroplast, and mitochondrial data; **(D)** Huang et al. (2019): MP/BI *analyses* based on chloroplast and nuclear data. Numeric values besides each genus correspond to the total number of accessions sampled per genus and the number below each figure represents the total number (individual/species) of Urticeae samples used in each respective analysis. "*" indicates full support; "–" indicates no support in **(A,C)**, support values of < 75 (MP) and < 0.95 (BI) in **(B)**, and support values of < 50% (MP) and < 0.7 (BI) in **(D)**.

silica gel-dried leaves and herbarium samples. Genomic DNA from each sample was then assessed for quality and quantity using both a NanoDrop 2,000 spectrophotometer (Thermo Fisher Scientific, United States) and agarose gel electrophoresis before library preparation. The library was built using the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England BioLabs) according to the manufacturer's instructions. Sequencing was then done using the Illumina HiSeq X Ten platform, yielding 150 bp paired-end reads. For each individual, 2–4 Gb of clean data was generated.

Assembly and Annotation

SPAdes (Bankevich et al., 2012) was used for *de novo* assembly of all sequences using kmer length of 85–111 bp. For the CP genome, we visualized and filtered the newly assembled contigs to generate a complete circular genome in both Bandage v. 0.80 (Wick et al., 2015) and Geneious v. 8.1 (Kearse et al., 2012). The newly assembled sequences were annotated using the reference genome *Debregeasia longifolia*_MBD01 (MN18994) in the Plant Genome Annotation (PGA) platform (Qu et al., 2019), followed by manual curation of genes in Geneious to check if the start and stop codons are correct. Furthermore, for CP genomes, tRNAscan-SE v. 1.21 (Schattner et al., 2005) was used to further verify the tRNA genes with default settings. We used Chloroplot (Zheng et al., 2020) to generate the physical maps of the CP genomes.

Plastome Structural Variation Analyses

Patterns of Inverted Repeat Boundary Shifts and Inversion

We characterized the genomic features of the 57 unique plastomes, including their size, structure (SC and IR regions), protein coding (PCG) and other (tRNA and rRNA) genes, and GC content. The junctions between the IR and single copy (SC) regions were then compared and analyzed using Geneious v. 8.1 (Kearse et al., 2012). ProgressiveMAUVE (Darling et al., 2010) was used to detect gene rearrangements and inversions among Urticeae taxa with *Elatostema parvum* as the reference genome. Default settings were used in ProgressiveMAUVE to automatically calculate the seed weight (15), and calculate Locally Collinear Blocks (LCBs) with a minimum LCB score of 30,000.

Repeat Sequence Analyses

We searched for the occurrence and distribution of three types of repeats within the studied plastomes of the tribe Urticeae. First, the program REPuter (Kurtz et al., 2001) was used to identify dispersed repeat sequences (forward, reverse, complement, and palindromic) using the following constraint values: a hamming distance of 3, minimum repeat size of 30 bp, and a maximum computed repeat of 100. Second, the tandem repeats were identified using the online program Tandem Repeats Finder (Benson, 1999) with the alignment parameters match, mismatch, and indels set to 2, 7, and 7, respectively. For this analysis, the maximum period size and TR array size were limited to 500 and 2,000,000 bp, respectively, and the minimum alignment score for reporting repeats was set at 50. Third, we used a Perl-based microsatellite identification tool (MISA; Thiel et al., 2003) to search for simple sequence repeats (SSRs) (i.e., mono-, di-, tri-, tetra-, penta-, and hexanucleotide repeats) within Urticeae plastomes. The threshold values for this analysis were set at 10, 6, 5, 5, 5, and 5 for mono-, di-, tri-, tetra-, penta- and hexanucleotides, respectively.

Sequence Divergence Analyses

To illustrate interspecific sequence variation and gene organization of the entire plastomes across the 57 examined species, we used mVISTA with the shuffle-LAGAN mode (Frazer et al., 2004) and *E. parvum* as the reference genome. For the assessment of sequence divergence and exploration of highly variable chloroplast markers, a sliding window analysis

was performed in DnaSP v. 6 (Rozas et al., 2017) to compute the nucleotide diversity (π) for all protein-coding (CDS) and non-coding (nCDS i.e., intron and intergenic spacer) regions. The step size was set to 300 bp, with a window length of 1,000 bp. The gene recovered to have the highest nucleotide diversity was then used to draw a phylogenetic tree to test the resolution of the identified barcode for our species.

Phylogenetic Inference

Phylogenetic analyses were conducted using different partitioning schemes from two datasets: the genome skimming [CP genome and the 18S-ITS1-5.8S-ITS2-26S (nrDNA) sequences] and two-locus (ITS and the *trnL-F* intergenic spacer) dataset. We extracted the coding (CDS) and non-coding (nCDS) regions from the CP genome to elucidate the phylogenetic utility of the different regions. This partitioning is important as both CDS and nCDS regions have been shown to exhibit distinct rates of nucleotide substitution (Wolfe et al., 1987; Jansen and Ruhlman, 2012). In total, six molecular data matrices were generated to explore the phylogenetic relationships of the tribe Urticeae, of which five were from the genome skimming dataset: (1) Whole chloroplast (CP) genomes, (2) CP coding regions (CDS), (3) CP non-coding regions (nCDS), (4) nuclear ribosomal DNA (nrDNA), and (5) combined whole CP genomes and nuclear ribosomal DNA (CP + nrDNA). The final matrix (6) sampled the two-locus dataset *trnL-F* intergenic spacer and ITS sequences (*trnL-F* + ITS) across expanded taxonomic sampling of 291 accessions.

Phylogenetic analyses were conducted using maximum likelihood (ML) and Bayesian inference (BI) methods in RAxML v. 8.2.11 (Stamatakis, 2014) and MrBayes v. 3.2 (Ronquist et al., 2012), respectively. Substitution models for all the datasets were first determined based on Akaike information criterion (AIC; Akaike, 1973) in jModelTest2 v. 2.1.7 (Darriba et al., 2012; **Supplementary Table 2**). Maximum likelihood analyses was done in RAxML using the bootstrap option of 1,000 replicates. For BI analyses, we performed two independent runs, each consisting of four Markov Chain Monte Carlo (MCMC) chains, and sampling of one tree every 1,000 generations for 1 million (CP, nCDS, and CP + nrDNA), 3 million (CDS), and 20 million (*trnL-F* + ITS and only nrDNA) generations. The convergence of the MCMC chains of each run was determined when the average standard deviation of split frequencies (ASDSF) achieved ≤ 0.01, and adequate mixing was based on the Effective Sample Sizes (ESS) values ≥ 200. Stationarity was assessed by checking if the plot of log-likelihood scores had plateaued in Tracer v1.7.1 (Rambaut et al., 2018). The first 25% of the sampled trees acquired from all the runs were discarded as burn-in, and consensus trees were constructed from the remaining trees to estimate posterior probabilities.

RESULTS

Chloroplast Genome Organization

The plastomes of Urticeae species varied greatly in sequence length, ranging in size from 145,419 bp (*Nanocnide lobata*)

TABLE 1 | Summary of sizes of the whole Urticeae plastomes, and the three compartments.

Species	Nucleotide length (bp)			
	Genome	LSC	SSC	IR
*Dendrocnide basirotunda*_J2078	152,646	83,433	18,229	25,492
*Dendrocnide meyenia*_D7	152,621	83,430	18,149	25,521
*Dendrocnide sinuata*_J7885	152,559	83,348	18,187	25,512
*Dendrocnide urentissima*_D4	152,658	83,444	18,230	25,492
*Discocnide mexicana*_W268	153,327	83,841	17,580	25,953
*Girardinia bullosa*_A1	152,388	82,974	17,728	25,833
*Girardinia chingiana*_G1	152,659	83,451	18,068	25,570
*Girardinia diversifolia*_G56	152,979	83,636	18,127	25,608
*Girardinia formosana hayata*_G3	152,596	83,364	18,056	25,588
Girardinia suborbiculata subsp. *grammata*_G22	152,687	83,453	18,020	25,607
Girardinia suborbiculata subsp. *suborbiculata*_G15	152,894	83,650	18,104	25,570
Girardinia suborbiculata subsp. *triloba*_G19	152,874	83,516	18,142	25,608
*Hesperocnide tenella*_W61	146,864	79,555	17,691	24,809
*Laportea aestuans*_L30	153,521	82,883	16,500	27,609
*Laportea bulbifera*_GLGE14842	149,436	81,759	17,859	24,909
*Laportea canadensis*_W167	150,253	82,394	17,783	25,038
*Laportea cuspidata*_L27	149,149	80,905	17,450	25,397
*Laportea decumana*_L15	151,855	82,777	18,080	25,499
*Laportea grossa*_L2	161,930	83,658	19,838	29,217
*Laportea medogensis*_GLGE141037	150,196	82,385	17,759	25,026
*Laportea mooreana*_L12	150,827	81,878	18,371	25,289
*Laportea ovalifolia*_L14	153,659	82,193	16,596	27,435
*Nanocnide japonica*_N3	145,970	78,396	17,300	25,137
*Nanocnide lobata*_N6	145,419	77,955	17,258	25,103
*Obetia aldabrensis*_W291	153,239	84,219	18,628	25,196
*Poikilospermum cordifolium*_Poi7	153,801	84,436	18,617	25,374
*Poikilospermum lanceolatum*_Poi8	153,879	84,521	18,618	25,370
*Poikilospermum naucleiflorum*_Poi6	153,782	84,414	18,600	25,384
*Touchardia latifolia*_T2	152,871	84,003	18,252	25,308
*Urera baccifera*_Ur21	153,215	84,314	18,027	25,437
*Urera cameroonensis*_Ur12	153,212	83,990	18,532	25,345
*Urera capitata*_W143	153,771	84,297	18,626	25,424
Urera cf *cordifolia*_Ur15	153,214	83,992	18,536	25,343
*Urera glabra*_Ur17	152,663	83,499	18,502	25,331
*Urera hypselodendron*_Ur16	153,212	84,007	18,515	25,345
*Urera oligoloba*_Ur23	153,919	84,056	18,561	25,151
*Urera robusta*_Ur19	153,198	84,017	18,491	25,345
*Urtica angustifolia*_J3303	146,703	79,830	17,683	24,595

(Continued)

TABLE 1 | (Continued)

Species	Nucleotide length (bp)			
	Genome	LSC	SSC	IR
*Urtica ardens*_GLGE152058	146,795	79,693	17,686	24,708
*Urtica atrichocaulis*_S11193	146,717	79,884	17,633	24,600
*Urtica chamaedryoides*_W162	146,455	79,304	17,701	24,725
Urtica dioica subsp. *xijiangensis*_U41	147,935	79,627	17,530	25,389
*Urtica dioica*_W174	146,928	80,052	17,676	24,600
*Urtica domingensis*_W145	146,125	79,260	17,665	24,600
*Urtica hyperborea*_J5455	147,898	79,748	17,588	25,281
*Urtica kioviensis*_U24	146,725	79,855	17,666	24,602
*Urtica macrorrhiza*_U50	146,747	79,886	17,661	24,600
*Urtica magellanica*_U33	146,606	79,613	17,657	24,668
*Urtica mairei*_J1664	146,790	79,689	17,685	24,708
*Urtica membranifolia*_S13031	158,078	79,719	17,689	30,335
*Urtica morifolia*_U200	146,755	79,643	17,690	24,711
*Urtica radicans*_U21	146,667	79,819	17,662	24,593
*Urtica rupestris*_U28	146,751	79,859	17,696	24,601
Urtica sp_U19	147,508	79,069	17,669	25,385
*Urtica thunbergiana*_J2498	146,846	79,667	17,711	24,734
*Urtica urens*_W175	147,516	79,076	17,668	25,386
*Zhengyia shennongensis*_Zh1	150,109	81,186	17,885	25,519

LSC, Large Single Copy; SSC, Small Single Copy; IR, Inverted Repeat.

to 161,930 bp (*Laportea grossa*) (**Table 1**). All exhibited a quadripartite structure typical of angiosperms (**Figure 2A**)—a pair of IRs (24,593–30,335 bp) separated by the LSC (77,955–84,521 bp) and SSC regions (16,500–19,838 bp). We observed a marginal difference in the GC content across the whole plastome (36.3–37.2%) and its elements — the IR (41.8–43.3%), LSC (33.8–34.7%), and SSC (29.6–31.1%) regions.

A range of 110–112 unique genes was detected across these plastomes, including 76–78 PCGs, 30 tRNA genes, and 4 rRNA genes. The IR region had complete duplications for 7 tRNA genes, 6 PCGs, and 4 rRNA genes. Across all 57 plastomes, 15 genes had a single intron *(atpF, ndhA, ndhB, petB*, *petD*, *rpl2*, *rpl16*, *rpoC1*, *rps16*, *trnA-UGC, trnG-UCC*, *trnI-GAU*, *trnK-UUU*, *trnL-UAA*, and *trnV-UAC*), while two genes (*clpP* and *ycf3*) had two introns. The *rps12* gene was spliced into two transcriptions, with one exon in the LSC and two in the IR. Notably, the *rpl2* gene of *Hesperocnide tenella* and most *Urtica* taxa except for *Urtica dioica* subsp. *xijiangensis*_U41, *Urtica dioica*_J5488, *Urtica hyperborea*_J5455, *Urtica* sp_U19, and *Urtica urens* lacked an intron. Apart from the region containing an inverted *trnN-GUU* in five species (four *Dendrocnide* species and *Laportea decumana*; **Figures 2B,C**), no significant gene rearrangement was observed within the studied plastomes (**Supplementary Figure 1A**).

Inverted Repeat Expansion and Contraction

Comparison of the IR boundaries among the 57 plastomes from tribe Urticeae revealed varying expansion and contraction of the IRs (**Figure 3A**). Herein, we report only the functional genes located at the IR-SC boundaries. The LSC/IRb border was embedded in the *rps19* gene (with 50–131 bp located within IRb) in 43 taxa. The remaining 14 species showed: an expansion in three species (*rpl22* in the LSC—*rps19* in the IRb); contraction (*rps*19 in the LSC—*rpl2* in the IRb) of the IR in three species; the loss of the *rps19* gene in eight species (*rpl22* in the LSC—*rpl2* in the IRb), causing variations in the boundary (**Figure 3B**). The IRb/SSC boundary generally fell within the *ndhF* gene (with 50–131 bp located at IRb), except in six species where the boundary was detected in the intergenic region of *trnNGUU-ndhF* (**Figure 3B**). We observed that the IRa/LSC boundary of most species lay within either the intergenic *rpl2-trnHGUG* or non-coding *trnH-GUG* regions, except for four species (*Hesperocnide tenella*_W61, *Urtica chamaedryoides*_W162, *Urtica magellanica*_U33, and *Urtica morifolia*_U200) in which the boundary was located within the intergenic region *trnH-GUG—psbA* (**Figure 3B**). The most conserved boundary across species was that of the SSC/IRa, which was always positioned within the *ycf1* coding gene, which had a length of 195–3,054 bp overlapping into the IRa region (**Figure 3B**).

Repeat Structure and Search for Simple Sequence Repeats

The 57 Urticeae plastomes showed a total of 6,274 repeats based on four classifications (**Figure 4A** and **Supplementary Table 3**). Generally, the most frequent repeat type was the SSR (2,919, 46.53%), followed by tandem (1,185, 18.89%), dispersed (1,140, 18.17%), and palindromic repeats (1,030, 16.42%) (**Figure 4A**). The distribution of the dispersed, tandem, and palindromic repeats varied between 25 (*Nanocnide japonica*_N3) and 124 (*Discocnide mexicana*_W268 and *Zhengyia shennongensis*_Zh1) (**Figure 4B**), and that of the number of SSRs ranged from 18 (*Laportea cuspidata*_L27) to 82 (*Laportea grossa*_L2) (**Figure 4C**). The majority of the SSRs were mononucleotides (2,627, 89.97%), with poly-A and poly-T SSR motifs being the two most frequent (**Figure 4D** and **Supplementary Table 5**). Dinucleotides, trinucleotides, tetranucleotides, pentanucleotides, and hexanucleotides accounted for 8.50, 1.27, 0.14, 0.03, and 0.07% of the SSR repeats, respectively (**Figure 4C** and **Supplementary Table 4**).

Sequence Divergence Analysis

Pairwise comparison of divergent regions within the 57 Urticeae plastomes using mVISTA revealed very low intra- and inter-generic (**Supplementary Figure 1B**) sequence divergence across the plastomes. Moreover, nCDS regions were generally more divergent and had higher levels of variation than CDS regions (**Supplementary Figures 1B, 2**). For the CDS, the top five genes with the highest nucleotide diversity (π) values (all with $\pi > 5\%$) were *rpoc2*, *cemA*, *rpoA*, *rpl22*, *ccsA*, and *ycf1*

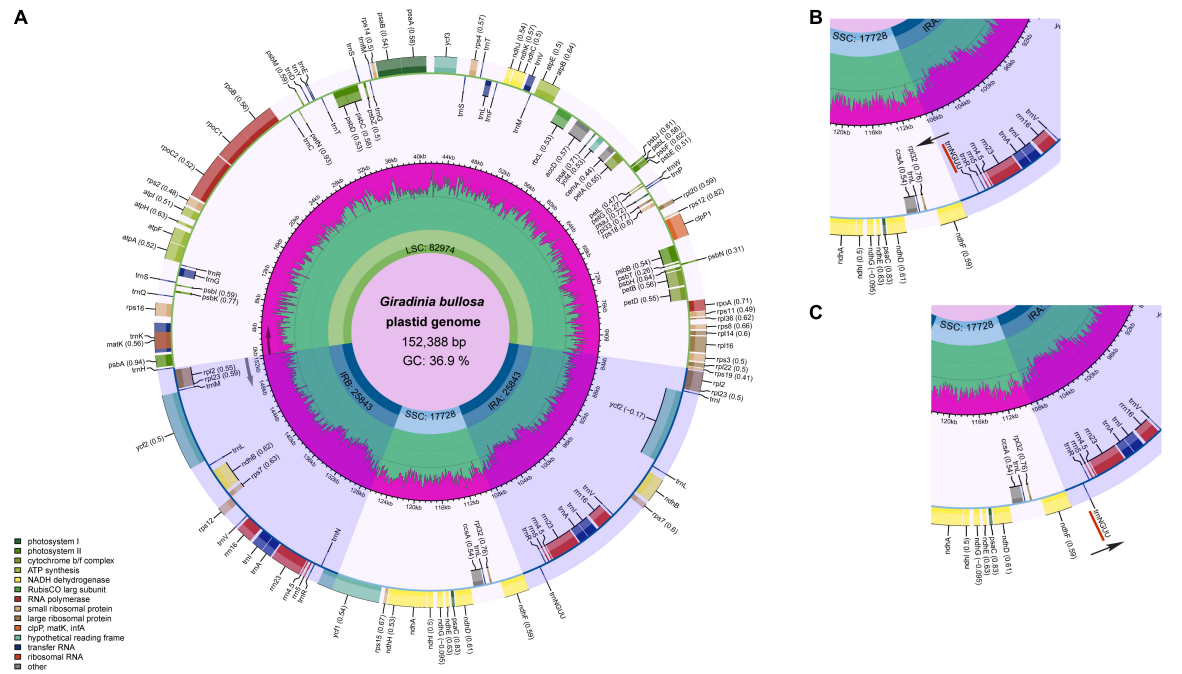

FIGURE 2 | (A) Gene map of complete chloroplast genome of *Girardinia bullosa* (a typical representative of gene organization in Urticeae *s.l.* plastomes); **(B)** inset map showing the inverted orientation of *trnN-GUU* in clade 3C except for *Discocnide mexicana*; **(C)** inset map of the Urticeae plastome, showing the typical orientation of trnN-GUU. Genes inside and outside the outer circle are transcribed clockwise and counterclockwise, respectively.

(**Supplementary Figure 2A**). The most variable nCDS regions were the *trnQ(UUG)—psbK*, *trnG(GCC)—trnfM(CAU)*, *ycf3—trnS(GGA)*, *cemA—petA*, and *ndhE—ndhG* spacer regions, all with $\pi > 10\%$ (**Supplementary Figure 2B**). The *ycf*1 gene tree depicted highly resolved and supported relationships, owing to the gene's high nucleotide diversity (**Supplementary Figure 2C**).

Phylogenetic Relationships

The sequence characteristics, tree diagnostic values, and the best-fit model determined by jModelTest for all datasets are given in **Supplementary Table 2**. The phylogenetic results presented here are based on both ML and BI analyses. The ML and BI analyses generated here generally had nearly identical topologies with few differences at the shallow nodes. Factors driving discrepancies between the ML and BI topologies have been previously reported (Huelsenbeck, 1995; Sullivan and Joyce, 2005; Som, 2014). Of those, the optimality criterion and specific hypotheses in the modeling of sequence evolution are parsimonious to explain the few discrepancies between the ML and BI topologies inferred from the same data matrix in our study. In most cases, the phylogenetic relationships inferred from ML were discussed because it has the most supporting shreds of evidence from the morphological affinities between the known species within the tribe Urticeae. The phylogenetic relationships constructed for each data matrix are further reported.

Chloroplast Data Analyses

The CDS, nCDS, and whole CP phylogenetic trees were largely identical in their topologies with only a few exceptions concerning the relationships of two clades 3F3I and 3F3II (**Supplementary Figures 3A–CI**). In the CDS data, these were sister to one another, hence formed a monophyletic clade 3F3 (**Supplementary Figure 3A**). However, in the whole CP dataset, 3F3I was sister to both 3F3II, and 3F4, while in nCDS dataset, 3F3II was sister to both 3F3I and 3F4 (**Supplementary Figures 3B,CI**). Nevertheless, it should be noted that the whole CP dataset generally had better support compared to both the CDS and nCDS datasets.

nrDNA Data Analysis

Regarding relationships between major clades in Urticeae, the results from the nrDNA dataset (**Supplementary Figure 3CII**) recovered almost congruent relationships with that of the whole CP dataset (**Supplementary Figure 3CI**), other than a few discrepancies in particular major clades and phylogenetic placement of some species. For instance, in the nrDNA phylogeny, clade 3D (*Girardinia*) was recovered as sister to clade 3C (**Supplementary Figure 3CII**), whereas in whole CP phylogeny, clade 3D was recovered as sister to a clade comprising subclades 3C, 3B, and 3A (**Supplementary Figure 3CI**). The sister relationships of clade 3G, and those within clade 3E-F also changed depending on the dataset examined. Moreover, we found slight differences in some shallower relationships between the whole CP and nrDNA phylogenies (e.g., the contradicting phylogenetic positions of *Dendrocnide urentissima*, *Girardinia suborbiculata* subsp. *suborbiculata*, etc.; **Supplementary Figure 3C**). These differences were, however,

mostly restricted to areas of poor support, and the whole CP phylogeny was generally better supported than that of nrDNA.

Combined Whole Chloroplast Genome and nrDNA (CP + nrDNA) Analysis

Phylogenetic resolution and node support values were significantly improved by the combination of whole CP genome and nrDNA data (**Figure 5**). The phylogeny inferred from the combined data matrix was the best resolved and supported phylogenetic tree amongst all the other data matrices, and was more similar in topology to the three chloroplast data matrices (whole CP, CDS, and nCDS, regions) than the nrDNA one (**Figure 5** and **Supplementary Figures 3A–C**). The monophyly of Urticeae was strongly supported (BS/PP = 100/1), with Elatostemeae as its sister tribe (**Figure 5**). Generally, the phylogeny was well resolved, with most nodes being strongly supported by both ML and BI analyses, except the placement of *Zhengyia shennongensis* (BS = 100 PP = "–"), the relationship between *Urtica domingensis* and *Hesperocnide tenella* (BS = "–" PP = 1), and the relationship between *Laportea aestuans* and *Laportea ovalifolia* (BS = "–" PP = 1) (**Figure 5**). Nine genera within Urticeae were recovered as monophyletic (*Dendrocnide*, *Discocnide*, *Girardinia*, *Hesperocnide, Obetia*, *Nanocnide*, *Poikilospermum*, *Touchardia,* and *Zhengyia*) and three as polyphyletic (*Urtica*, *Laportea*, and *Urera*), all with strong support. For ease of discussion, we sectioned Urticeae into six major clades, each with full bootstrap support; the names reflect the clade naming system of Wu et al. (2013). They include Clade 3A (*Urtica*, *Hesperocnide,* and *Zhengyia*), Clade 3B (*Nanocnide* and *Laportea cuspidata*), Clade 3C (*Dendrocnide*, *Discocnide,* and *Laportea decumana*), Clade 3D (*Girardinia*), and Clade 3G (*Laportea*). Clade 3E-F was recovered as sister to the rest of the Urticeae tribe with maximum support, and comprised *Poikilospermum*, *Urera*, *Obetia*, and *Laportea*. Within it, *Poikilospermum* (sub-clade 3F4) was recovered for the first time as a sister clade to *Urera* (sub-clade 3F3) with full support (**Figure 5**). *Urera* comprised three separate subclades within Clade 3E-F, each with strong support. Moreover, in this study *Laportea* was split into five different clades. Clade 3D (*Girardinia*) was also recovered for the first time as sister to a clade comprising 3A, 3B, and 3C, with full support.

Combined Analysis of *trnL-F* + ITS

The tree topology from the analysis of the *trnL-F* and ITS dataset was largely congruent with the previously published phylogenies inferred from a small number of loci. Eight genera were strongly

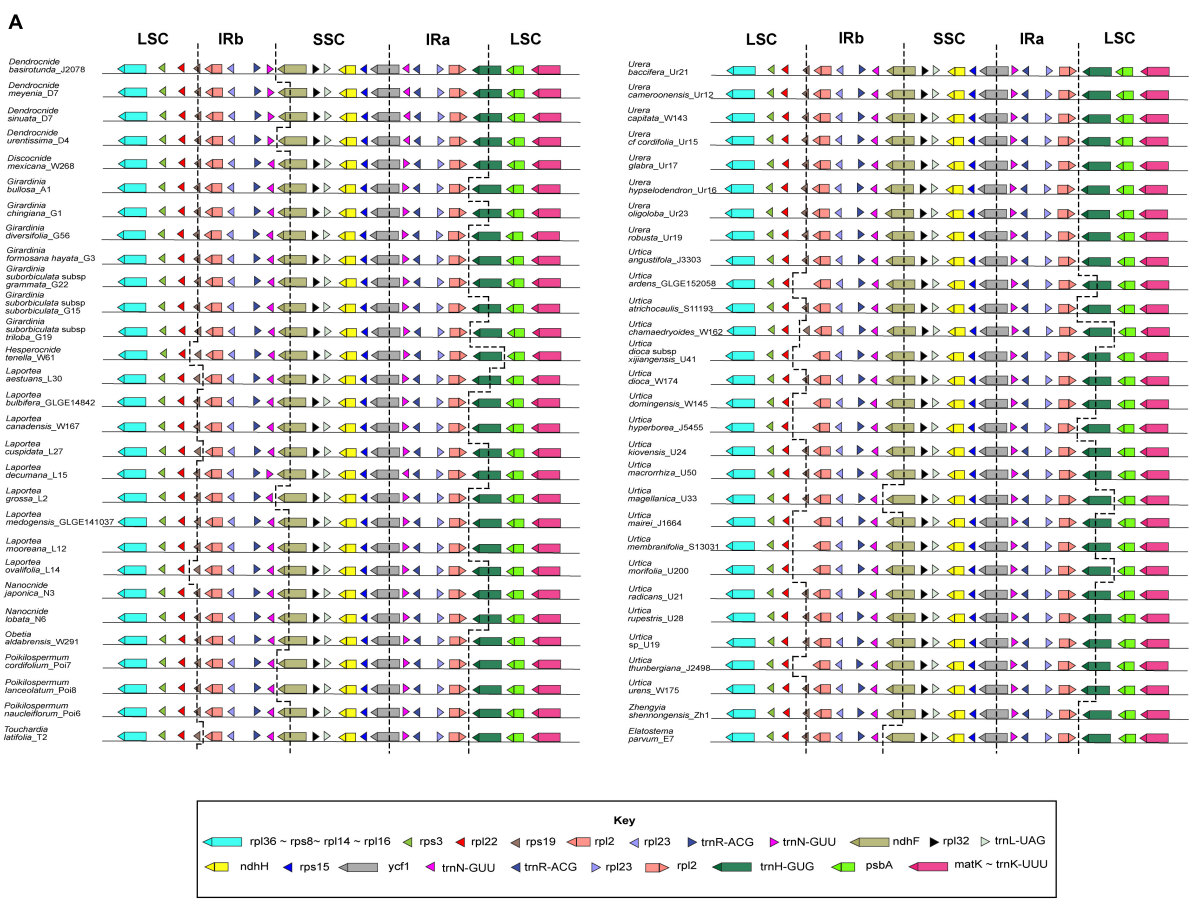

FIGURE 3 | (Continued)

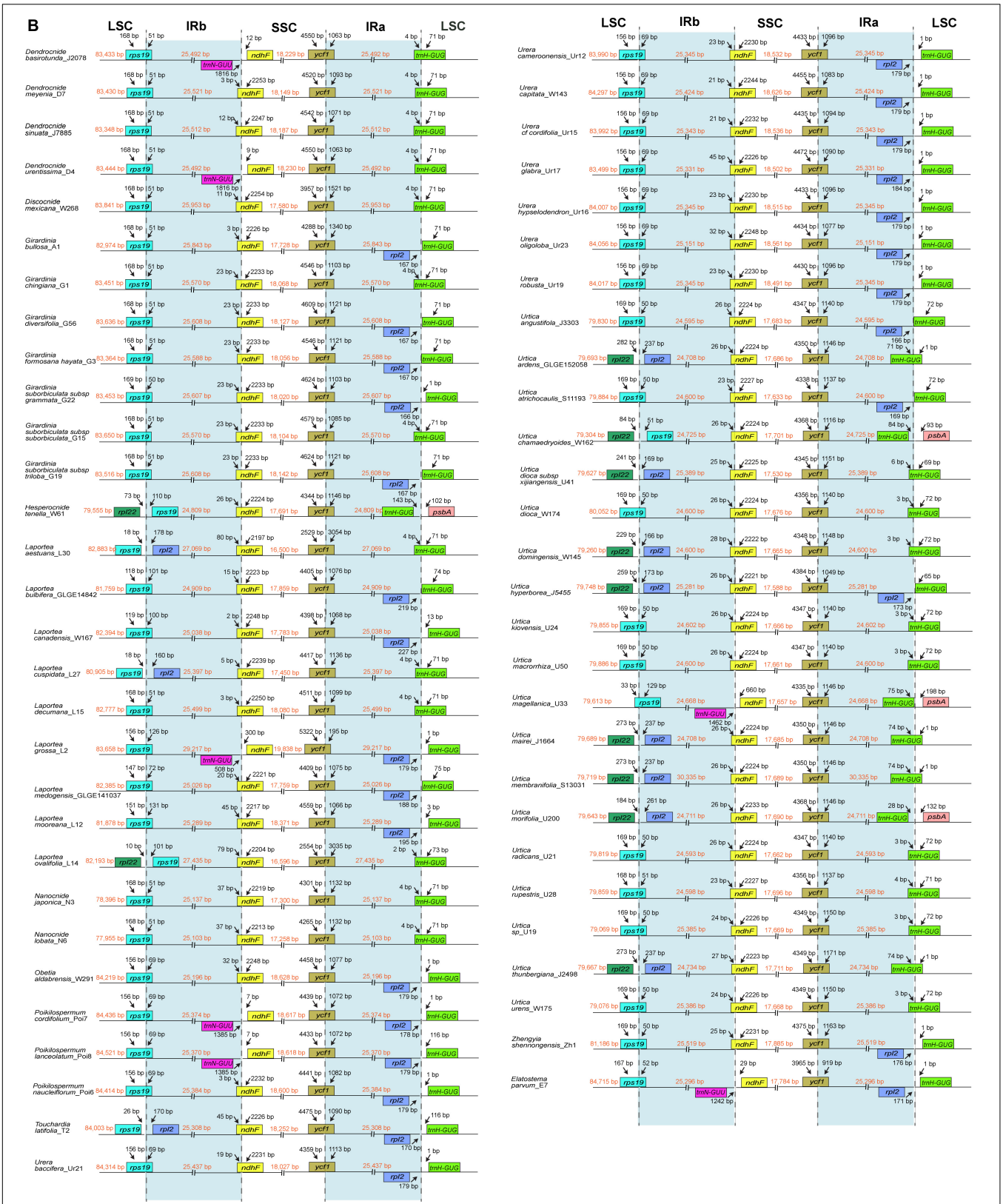

FIGURE 3 | (A) representative map showing expansions and contractions in the IR region; **(B)** comparison of the IR/SC junctions among 57 Urticeae plastomes. The genes around the borders are shown above or below the main line. LSC, Large Single Copy; SSC, Small Single Copy; IR (a and b), Inverted Repeat a and b.

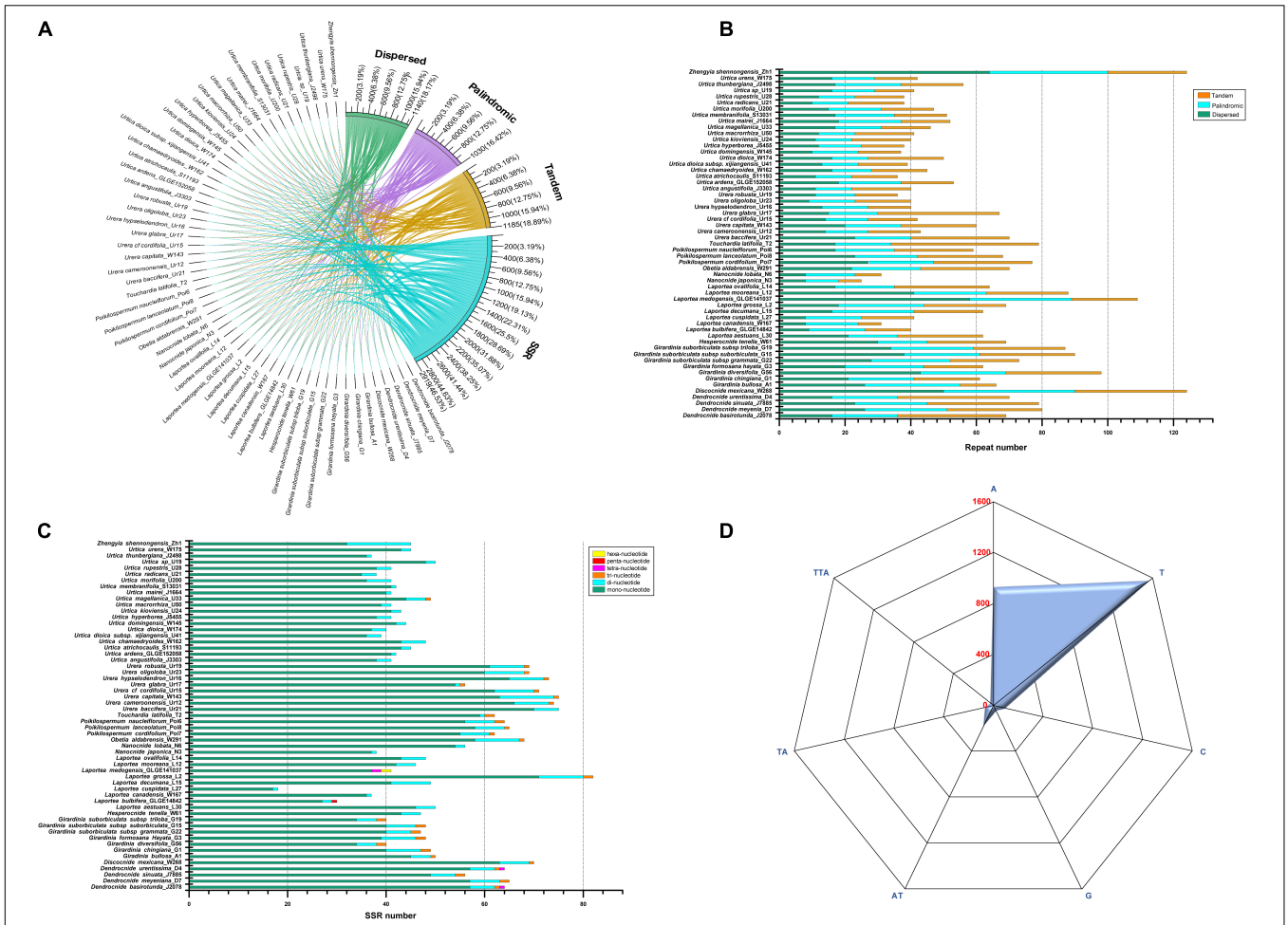

FIGURE 4 | (A) Chord diagram showing a connection between species and their corresponding repeat types (Dispersed, Palindromic, Tandem, and SSR). The tick marks beside each repeat type indicate the frequency of the number of repeats detected and their percentages, respectively; **(B)** frequency of tandem, palindromic, and dispersed repeats; **(C)** number of the six SSR(simple sequence repeats) nucleotides; **(D)** the seven most abundant SSR motifs in Urticeae plastomes. The values along the axis represent the abundance values.

supported as monophyletic (i.e., *Dendrocnide*, *Discocnide*, *Girardinia*, *Obetia*, *Nanocnide*, *Poikilospermum*, *Touchardia,* and *Zhengyia*) while four genera were recovered as polyphyletic (i.e., *Hesperocnide*, *Urtica*, *Laportea*, and *Urera). Hesperocnide* was recovered here as polyphyletic (BS/PP > 90/0.90 and BS/PP < 90/0.90; **Figure 6**) as compared to the combined whole (CP + nrDNA) where it was retrieved as monophyletic with full bootstrap support (**Figure 5**). Moreover, most of the shallow nodes of *trnL-F* and *ITS* tree received lower bootstrap support (**Figure 6**) compared to the combined whole (CP + nrDNA) tree, in which nearly all the nodes were fully supported.

DISCUSSION

Plastome Structural Evolution

All 57 Urticeae CP genomes examined are quadripartite but varied in size. The observed range was consistent with chloroplast genome sizes of angiosperms (Zhang et al., 2021) and the few existing sequenced plastomes of Urticaceae (Wang et al., 2020b; Li et al., 2021), which range between 120 and 180 kb. Of the plastomes in our study, *Laportea grossa* had the largest genome, while *Nanocnide lobata* had the smallest, implying that CP genomes in Urticaceae are structurally different. Also, the number of PCGs in the Urticeae plastomes in our study (76–78) was comparable with the typical range for angiosperm plastomes (70–88 genes) (Wicke et al., 2011). Likewise, we found congruence with the range of GC content previously reported in other plastomes of Urticaceae, e.g., *Pilea mollis* (36.72%; Li et al., 2021), *Elatostema dissectum* (36.2%; Fu et al., 2019), *Droguetia iners* (36.9%), and *Debregeasia elliptica* (36.4%) (Wang et al., 2020b). Generally, the GC content had no significant phylogenetic implication in our study. Moreover, consistent with previous studies (Li et al., 2020, 2021; Dong et al., 2021), the GC content was higher in the IR than in the SC. The GC inequality perhaps also plays a significant factor in the conservatism of the IR region compared to the SC regions (Li et al., 2020).

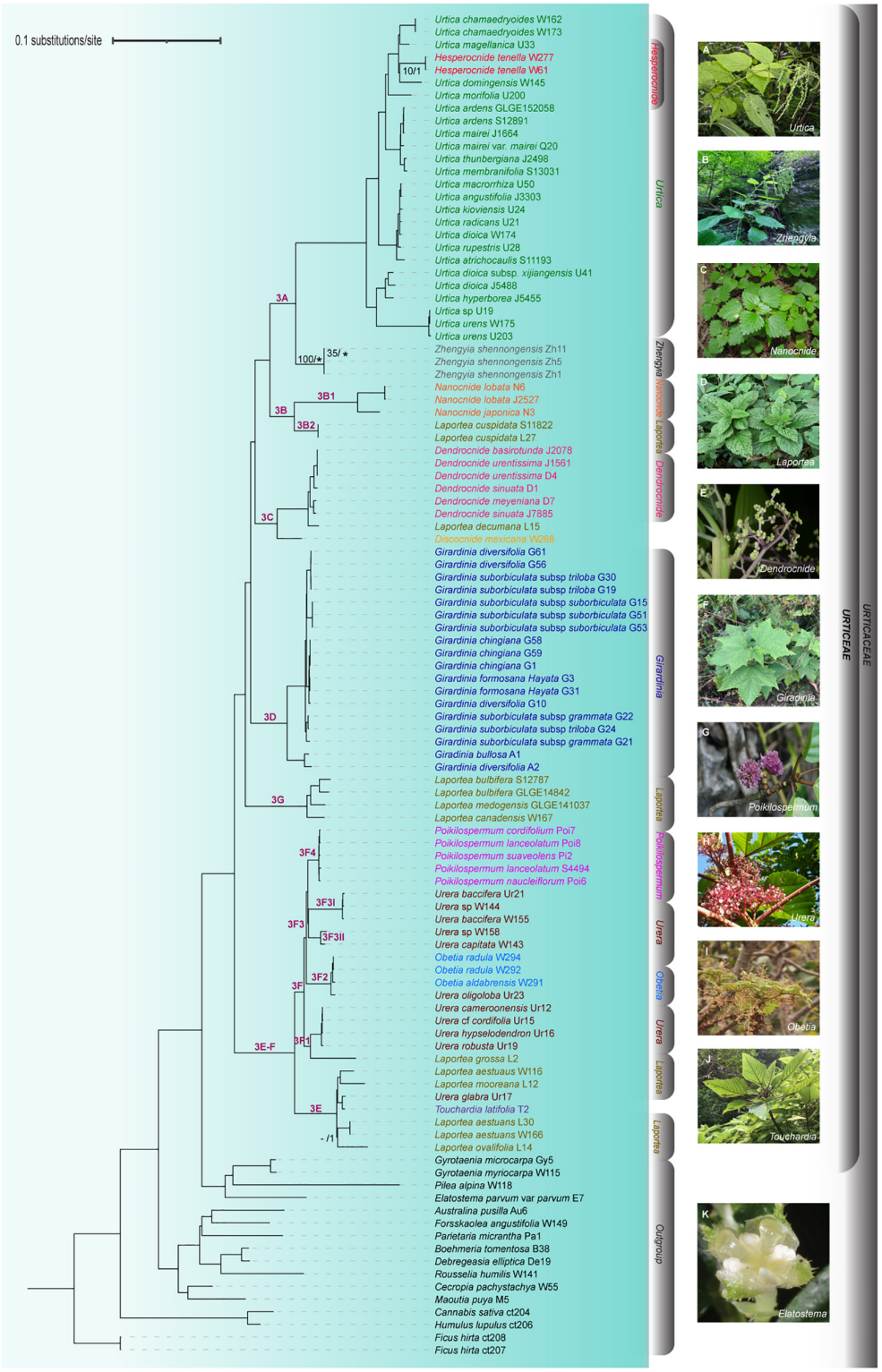

FIGURE 5 | Phylogenetic relationships of Urticeae inferred from maximum likelihood (ML) and Bayesian inference (BI) based on combined complete plastome and nrDNA sequences. Numbers on the branch indicate clade classification (in purple) and ML_BS/BI_PP values (in black)—note that branches with no support values indicate both ML_BS ≥ 90 and BI_PP = 1.00; lastly, "*" indicate incongruence between ML and BI trees and "–" no support values. Representative images of genera within Urticeae *s.l.* are shown on the right. Photographs: **(A–C,E,G,K)** by Z.Y. Wu, **(D,F)** by C.A. Ogoma, **(H)** by U. Dreschel, **(I)** by C. Kunath, and **(J)** photographed by J. Cantley.

Among the genes present in our Urticeae plastomes, *rpl2* was noteworthy, considering that 18 of the examined species had no introns for this gene. Intron loss has been widely documented in angiosperm plastomes: e.g., *Avena sativa* (*rpoC1* intron loss; Liu et al., 2020b), *Cicer arietinum* (*rps12* and *clpP* intron losses; Jansen et al., 2008), *Lagerstroemia* (*rpl2* intron loss; Gu et al., 2016), and Asteropeiaceae + Physenaceae (*rpl2* intron loss; Yao et al., 2019). Another notable structural change found here was an inversion of the *trnN-GUU* gene, which is a synapomorphy of the clade 3C, except for the clade's basal species *Discocnide mexicana* (**Figure 2B**). Gene inversions have also been detected in many angiosperm plastomes, including those of Poaceae (Guisinger et al., 2010), Styracaceae (Yan et al., 2018), Orchidaceae (*Uncifera acuminata*; Liu et al., 2020a), and Adoxaceae (Wang et al., 2020a). The latter, involving the inversion of the *ndh*F gene in Adoxaceae, is relevant to our study since it involves only one gene that also borders the inverted gene in our study (*trnN-GUU*). Typically, plastome inversions are deemed highly valuable in phylogenetics owing to their relative rarity, easily determined homology, and easily inferred state polarity (Cosner et al., 1997; Dugas et al., 2015; Schwarz et al., 2015). Despite some significant research efforts regarding the intramolecular recombination between dispersed short inverted/direct repeats and tRNA genes (Cosner et al., 1997; Haberle et al., 2008; Sloan et al., 2014), the cause of inversions in plant genomes remains unclear.

Our analyses showed that IR expansion and contraction vary across Urticeae, and lack taxonomic utility at a broader scale. Mostly, the SC/IR borders are relatively conserved among angiosperm plastomes and usually located within the *rps19* or *ycf1*gene (Downie and Jansen, 2015), even though it is assumed that IR expansion or contraction is accompanied by the shift of genes located in the IR/SC boundary (Zhu et al., 2016). Similar IR/SC changes are also evident in other Urticaceae plastomes (Wang et al., 2020b; Li et al., 2021). Changes in the IR/SC junctions have been considered one of the main drivers of the size diversity in the CP genomes of higher plants (Ma et al., 2013; Yang et al., 2016; Yan et al., 2018; Xue et al., 2019). Notably, we found the loss of the *rps19* gene to be the most parsimonious explanation for the diversification of the genes bordering the IR/LSC in the eight plastomes examined from the genus *Urtica*—(*U. ardens*_GLGE152058, *U. dioica* subsp. *xijiangensis*_U41, *U. domingensis*_W145, *U. hyperborea*_J5455, *U. mairei*_J1664, *U. membranifolia*_S13031, *U. morifolia*_U200, and *U. thunbergiana*_J2498; **Figure 3A**).

We detected several repeat types within the sampled plastomes of tribe Urticeae, among which SSRs were the most frequent, accounting for 46.53% of the repeats (**Figure 4A**). The most abundant SSRs were mononucleotide homopolymers, particularly poly−A and T motifs (**Figure 4D** and **Supplementary Table 5**). This phenomenon of A/T motif abundance has also been reported in *Pilea* (Li et al., 2021) and *Debregeasia* (Wang et al., 2020b) species, and might occur because the A/T motifs are more frequently dynamic compared to G/C (Li et al., 2020). Generally, it is presumed that repeat sequences are closely connected with a vast number of indels; therefore, the more abundant they are, the greater the nucleotide diversity (McDonald et al., 2011). Hence, the chloroplast repeat sequences could be potential sources of variation for evolutionary studies, and population genetics (Xue et al., 2012). We also found higher nucleotide diversity in the nCDS than in the CDS regions, consistent with findings from other taxa (Jansen and Ruhlman, 2012; Huang et al., 2014). Although the nucleotide content of chloroplast genomes is usually relatively stable, with a highly conserved gene structure

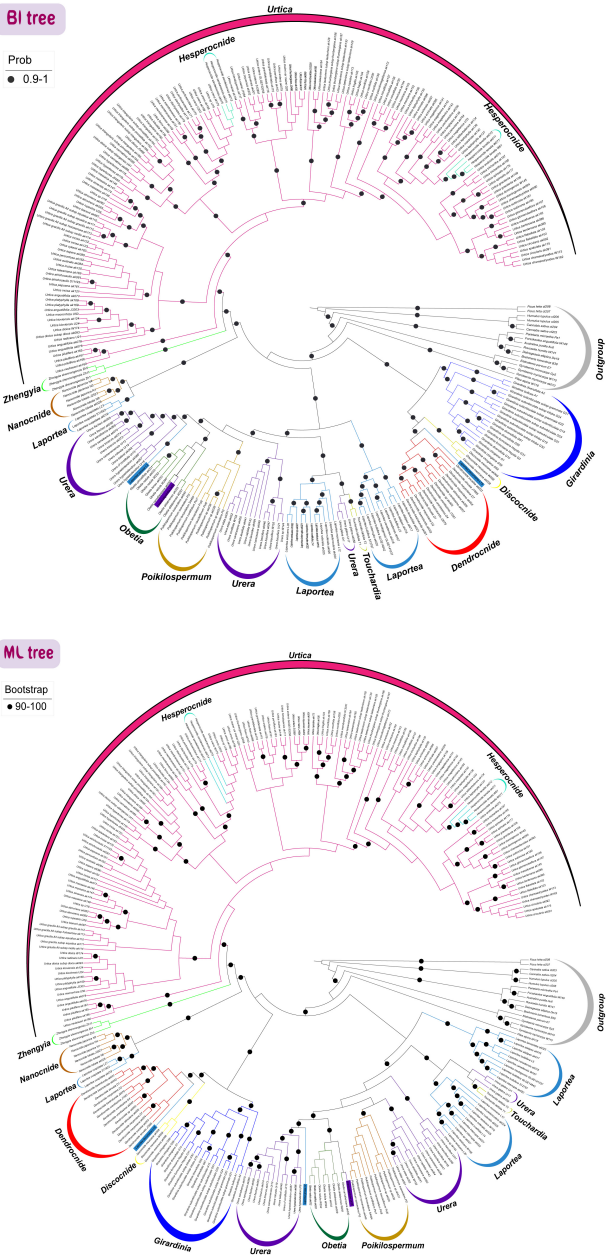

FIGURE 6 | Phylogenetic relationships of Urticeae tribe inferred from maximum likelihood (ML) and Bayesian inference (BI) based on *trnL-F* intergenic spacer and ITS (*trnL-F* + ITS) loci sequences. Support values indicated by black circle show ML_BS ≥ 90 and BI_PP ≥ 0.90.

(Jansen et al., 2005; Ravi et al., 2008; Wicke et al., 2011), mutation hotspots still exist within it (Zhang et al., 2021). We detected a total of 11 hypervariable loci in both CDS and nCDS regions (**Supplementary Figure 2**) that could be potentially used as DNA barcodes in future studies of this group. Among them was the locus *ycf1*, which was also reported in previous Urticaceae studies (Wang et al., 2020b; Li et al., 2021) as a highly variable locus with great taxonomic utility. Moreover, a study by Dong et al. (2015) reinforces this view, and recommemnds *ycf*1 as a suitable plastid barcode for land plants. Indeed, our *ycf*1 phylogenetic tree (**Supplementary Figure 2C**) is consistent with the above studies, especially with regard to the high resolution and support level. Therefore, we suggest that *ycf1* represents a highly useful molecular marker, not just for tribe Urticeae, but likely for the entire family. Presently, DNA barcodes are widely used in species identification, resource management, and studies of phylogeny and evolution (Gregory, 2005; Liu et al., 2019).

Phylogenetic Relationships of Urticeae

Phylogenetic Relationships Based on Genome Skimming (CP Genome + nrDNA) Data

The combined matrix (CP genome + nrDNA) yielded a well-supported phylogeny and resolved many relationships of the tribe Urticeae depite the topological difference in clades 3(D, 3G, and E-F), between the two separate datasets (**Supplementary Figure 3C**). This resolution shown by the combined matrix may be ascribed to the greater number of phylogenetically informative plastid sites (**Supplementary Table 2**). Moreover, it could be due to a weak phylogenetic signal in the nrDNA that agrees and complements the signal of the CP matrix. However, beyond some major conflicts, the individual CP and nrDNA trees are generally in agreement with most conflicting relationships pertaining to poorly supported areas of the phylogeny, although we did not perform follow-up analyses to identify what this means for different parts of the tree. Cases of topological dissimilarity are often reported in phylogenetic studies (Wendel and Doyle, 1998; reviewed by Degnan and Rosenberg, 2009). This phenomenon can be best explained by a number of factors including differences in taxon sampling, incomplete lineage sorting, hybridization/introgression, paralogy, gene duplication and/or loss, and horizontal gene transfer (Degnan and Rosenberg, 2006; Naciri and Linder, 2015; Lin et al., 2019; Nicola et al., 2019). Hence, as more samples become available, future studies should investigate the factors responsible for the observed conflicting relationships within the Urticeae.

Our study represents the first phylogeny of the tribe Urticeae based on a broad sampling of both CP genomes and nrDNA sequences. Importantly, we clarify which of the Urticeae genera are strongly supported as monophyletic or polyphyletic (**Figure 5**). Compared to previous studies based on a limited number of genes (Hadiah et al., 2008; Deng et al., 2013; Wu et al., 2013, 2018; Kim et al., 2015; Grosse-Veldmann et al., 2016; Huang et al., 2019; Wells et al., 2021), we exploited the utility of whole CP genomes for resolving phylogenetic relationships in Urticeae, and also revealed the most informative sites and regions across the plastome. Our results proved to be largely consistent with most of the recently established phylogenetic relationships of Urticeae based on a range of 3–7 selected marker regions (Wu et al., 2013, 2018; Kim et al., 2015; Huang et al., 2019; Wells et al., 2021). In general, however, our data improved resolution throughout Urticeae compared with previous studies, with almost all nodes being fully supported, especially those previously known to be problematic. Four of the most important new phylogenetic insights generated by the current study are discussed below.

First, the sister relationship of *Girardinia* has been contentious. *Girardinia* had been resolved as sister to *Dendrocnide-Discocnide* based on chloroplast, mitochondrial, and nuclear data (Wu et al., 2013), and using ITS, *rbcL*, and *trnL-F* regions (Kim et al., 2015), but without support in either case. Subsequently, using expanded taxon sampling and five markers from both the nuclear and CP genomes, the sister relationship of *Girardinia* to *Dendrocnide-Discocnide-Laportea-Nanocnide-Zhengyia-Urtica-Hesperocnide* was resolved, but with limited support (Wu et al., 2018; Huang et al., 2019). Our results support this latter relationship but with maximum support (BS/PP = 100/1), for the first time.

Second, our molecular phylogeny of the "*Urera* alliance clade" (this study clade 3E-F) corroborated the generic delimitation and subdivisions of the "*Urera* clade" from Wells et al. (2021), and showed two clades of *Laportea* (which they did not examine) as also a member (**Figure 5**). Their division of the paraphyletic *Urera* into three genera was strongly supported here: these were *Urera s.s.* (our Clade 3F3), *Scepocarpus* (entirely African; our clade 3F1, which also includes *Laportea grossa*), and an expanded *Touchardia* (part of clade 3E, that includes *Urera glabra* from Hawaii and three species of *Laportea* as per our study). Our data suggests that the two *Laportea* clades should hence be fully examined and considerations made as to whether to subsume them within the resurrected *Scepocarpus* and the expanded *Touchardia*.

Third, previous studies (Kim et al., 2015; Wu et al., 2018; Huang et al., 2019) have typically resolved *Laportea* into three clades. For instance, Kim et al. (2015) recovered three *Laportea* clades corresponding to sections *Laportea* Gaudich. (*L. alatipes*, *L. bulbifera*, *L. canadensis*, *L. lanceolata*), *Sceptrocnide* (Maxim.) C. J. Chen (*L. cuspidata*), and *Fleurya* (Gaudich.) Chew [*L. aestuans* (L.) Chew, *L. interrupta*, *L. ruderalis* (G. Forst.) Chew], consistent with the sectional classification of Wang and Chen (1995). Our analysis, however, resolved *Laportea* into five major clades. Moreover, we found that *L. aestuans* was polyphyletic: one subgroup was sister to *L. mooreana* with full support and the other was sister to *L. ovalifolia* with support of BS/PP = –/1. The latter relationship was detected by Wu et al. (2018) but without support. However, other studies found different relationships: *L. aestuans* as sister to *L. interrupta,* and *L. ruderalis* with full support according to Kim et al. (2015), or sister to *L. ruderalis* and *L. peduncularis* with support of MP/PP = 96/1 according to Huang et al. (2019). These discrepancies likely reflect differences in taxon and molecular sampling—with a wider sampling of populations, *L. aestuans* might comprise more than two unrelated clades. While additional study on

Laportea is clearly needed, the current study provides one of the most comprehensive phylogenetic perspectives on this little-studied genus. Future investigations should, however, employ more extensive molecular data across the entire phylogenetic spectrum of *Laportea* to further clarify its relationships and the number of lineages.

Finally, our analysis resolved the sister relationship between *Poikilospermum* and *Urera* previously obtained by Huang et al. (2019), but replacing their modest support (BS/PP = 65/0.89) with full support (BS/PP = 100/1) for the first time.

Comparison Between Genome Skimming (CP Genome + nrDNA) and Two-Locus (*trnL-F* + ITS) Phylogeny

In our study, the trees inferred from both the CP genome + DNA and the two-locus dataset (*trnL-F* + ITS) provided full support for the monophyly of Urticeae. However, the CP genome + nrDNA tree presented a higher percentage of fully supported nodes compared with that of the two-locus tree (**Figures 5**, **6**). This underscores the importance of genome-scale datasets for resolving major recalcitrant relationships.

The most notable finding from our two-locus phylogenetic analysis was the reconstruction of *Hesperocnide* as polyphyletic, consistent with Huang et al. (2019). Our current CP genome + nrDNA analysis and prior molecular studies, however, recovered *Hesperocnide* as monophyletic (Kim et al., 2015), with a close relationship to *Urtica* (Sytsma et al., 2002; Hadiah et al., 2008; Deng et al., 2013; Wu et al., 2013; Kim et al., 2015). The polyphyletic results from the two-locus tree can be ascribed to the sampling of members of the second species that were absent in the plastome analysis. Consequently, Wu et al. (2013) suggested that *Hesperocnide* be subsumed in the genus *Urtica*, since these two genera show some morphological similarities. However, owing to this equivocality about the phylogeny of *Hesperocnide*, we suggest a more rigorous examination of this genus to fully validate its status.

CONCLUSION AND FUTURE DIRECTIONS

Our study provides important novel insights on Urticeae phylogeny and plastome evolution. The detailed comparative analyses show that Urticeae plastomes exhibit striking differences in genome size, gene number, inversions, intron loss, sequence repeats, and IR/SC boundaries. These kinds of variations will be useful for studies on molecular marker discovery, population genetics, and phylogeny. Resolving the enigmatic relationships within tribe Urticeae has, to date, been a daunting task due to the paucity of genomic resources for the clade. Our study is the first to report phylogenetic relationships in Urticeae based on a broad sampling of whole plastome sequences. This dataset allowed for resolution of several recalcitrant branches (e.g., the relationship of *Poikilospermum* to *Urera,* the sister relationship of *Girardinia,* etc.) that were ambiguous in previous studies. Although our taxon sampling was sufficient to resolve relationships among the major clades in the tribe, additional sampling of particular genera (e.g., *Laportea*) and species (e.g., *Laportea aestuans* and *Hesperocnide sandwicensis*) would further refine our understanding of phylogenetic relationships in Urticeae. Building on the solid framework established here, future studies with even greater taxonomic and genomic sampling could contribute to a better understanding of the diversification patterns in Urticeae in relation to climatic, biogeographic, and ecological factors.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be accessed at NCBI GenBank; the list of accessions can be found in **Supplementary Table 1**.

AUTHOR CONTRIBUTIONS

Z-YW, D-ZL, JL, and CO conceptualized the study. Z-YW, JL, RM, AM, and YZ collected the samples. OO and CO conducted the analyses. CO and Z-YW drafted the manuscript. Z-YW, CO, GS, MW, OO, RM, D-ZL, and AM revised the manuscript. All authors read and approved the final manuscript.

FUNDING

Funding for this project was supported by the National Natural Science Foundations of China (31970356, 42171071, and 41971071), CAS' Youth Innovation Promotion Association (2019385), the Key Research Program of Frontier Sciences, CAS (ZDBS-LY-7001), the Top-notch Young Talents Projects of Yunnan Provincial "Ten Thousand Talents Program" (YNWR-QNBJ-2020-293 and YNWR-QNBJ-2018-146), and CAS Strategic Priority Research Program (XDB31000000). MW was supported by the Postdoctoral International Exchange Program of the Office of China Postdoctoral Council, and the Postdoctoral Targeted Funding and Postdoctoral Research Fund of Yunnan Province.

ACKNOWLEDGMENTS

We are really grateful to Qi Chen and Ruo-Nan Wang for their great assistance during the data analysis. We also immensely appreciate the following herbaria for providing access to study specimens: Kunming Institute of Botany (KUN), Royal Botanical Gardens, Kew (K), University of Florida Herbarium (FLAS), and Royal Botanical Gardens, Edinburgh (E). We thank the Royal Botanic Gardens Kew for providing some of the DNA materials. This work was facilitated by the Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences (CAS).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.870949/full#supplementary-material

Supplementary Figure 1 | (A) Mauve alignment showing gene arrangements within the studied 57 Urticeae plastomes (length indicated above). Large colored boxes represent the gene blocks and the colored lines indicates linear position of different genes in the plastome. **(B)** Comparison of 57 Urticeae CP genomes using mVISTA, with the *E. parvum* genome as the reference. The *y*-axis represents the percent identity within 50–100%. Gray arrows indicate the direction of gene transcription. Blue blocks indicate conserved genes, while red blocks indicate conserved non-coding sequences (CNS).

Supplementary Figure 2 | Variable sites in homologous regions of the 57 sampled plastomes from Urticeae. The *y*-axis represent the nucleotide diversity (Pi) of each window, and *x*-axis is the position of the midpoint of each window used in the Sliding window analysis. **(A)** Coding regions. **(B)** Non-coding regions. **(C)** The *ycf*1 gene tree depicting highly resolved and supported relationships achieved by the identified barcode.

Supplementary Figure 3 | (A) Phylogenetic relationships of Urticeae tribe inferred from maximum likelihood (ML) and Bayesian inference (BI) based on CP coding (CDS) regions. Support values above the branches are maximum likelihood bootstrap support values (ML_BS)/Bayesian posterior probabilities (BI_PP)—note: branches with no support values indicate both ML_BS $\geq$ 90 and BI_PP = 1.00—whereas "*" indicate incongruence between ML and BI trees. Major clades of Urticeae *s.l.* are indicated on the right, respectively. CDS, chloroplast coding region. **(B)** Phylogenetic relationships of Urticeae tribe inferred from maximum likelihood (ML) and Bayesian inference (BI) based on CP non-coding (nCDS) regions. Support values above the branches are maximum likelihood bootstrap support (ML_BS)/Bayesian posterior probability (BI_PP)—note that branches with no support values indicate both ML_BS $\geq$ 90 and BI_PP = 1.00—whereas "*" indicate incongruence between ML and BI trees. Major clades of Urticeae *s.l.* are indicated on the right, respectively. nCDS, chloroplast non-coding region. **(C)** Phylogenetic relationships of Urticeae tribe inferred from maximum likelihood (ML) and Bayesian inference (BI) based on integrated CP genome and nrDNA trees. Support values above the branches are maximum likelihood bootstrap support (ML_BS)/Bayesian posterior probability (BI_PP)—note that branches with no support values indicate both ML_BS $\geq$ 90 and BI_PP = 1.00—whereas "*" indicate incongruence between ML and BI trees. Major clades of Urticeae *s.l.* are indicated on the right, respectively. CP, Complete chloroplast genome; nrDNA, nuclear ribosomal DNA (18S-ITS1-5.8S-ITS2-26S).

REFERENCES

Akaike, H. (1973). "Information theory as an extension of the maximum likelihood principle," in *Second International Symposium on Information Theory*, eds B. N. Petrov and F. Csaki (Budapest: Akademiai Kiado), 267–281.

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

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580. doi: 10.1093/nar/27.2.573

Benvenutti, R. C., Dalla Vecchia, C. A., Locateli, G., Serpa, P. Z., Lutinski, J. A., Rodrigues Junior, S. A., et al. (2020). Gastroprotective activity of hydroalcoholic extract of the leaves of *Urera baccifera* in rodents. *J. Ethnopharmacol.* 250:112473. doi: 10.1016/j.jep.2019.112473

Bodros, E., and Baley, C. (2008). Study of the tensile properties of stinging nettle fibres (*Urtica dioica*). *Mater. Lett.* 62, 2143–2145. doi: 10.1016/j.matlet.2007.11.034

Conn, B. J., and Hadiah, J. T. (2009). Nomenclature of tribes within the Urticaceae. *Kew Bull.* 64, 349–352. doi: 10.1007/s12225-009-9108-4

Cosner, M. E., Jansen, R. K., Palmer, J. D., and Downie, S. R. (1997). The highly rearranged chloroplast genome of *Trachelium caeruleum* (Campanulaceae): multiple inversions, inverted repeat expansion and contraction, transposition, insertions/deletions, and several repeat families. *Curr. Genet.* 31, 419–429. doi: 10.1007/s002940050225

Darling, A. E., Mau, B., and Perna, N. T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* 5:e11147. doi: 10.1371/journal.pone.0011147

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

Degnan, J. H., and Rosenberg, N. A. (2006). Discordance of species trees with their most likely gene trees. *PLoS Genet.* 2:e68. doi: 10.1371/journal.pgen.0020068

Degnan, J. H., and Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* 24, 332–340. doi: 10.1016/j.tree.2009.01.009

Demenou, B. B., Migliore, J., Heuertz, M., Monthe, F. K., Ojeda, D. I., Wieringa, J. J., et al. (2020). Plastome phylogeography in two African rain forest legume trees reveals that Dahomey Gap populations originate from the Cameroon volcanic line. *Mol. Phylogenet. Evol.* 150:106854. doi: 10.1016/j.ympev.2020.106854

Deng, T., Kim, C., Zhang, D. G., Zhang, J., Li, Z. M., Nie, Z. L., et al. (2013). *Zhengyia shennongensis*: a new bulbiliferous genus and species of the nettle family (Urticaceae) from central China exhibiting parallel evolution of the bulbil trait. *Taxon* 62:89. doi: 10.1002/tax.621008

Di Virgilio, N., Papazoglou, E. G., Jankauskiene, Z., Di Lonardo, S., Praczyk, M., and Wielgusz, K. (2015). The potential of stinging nettle (*Urtica dioica* L.) as a crop with multiple uses. *Ind. Crops Prod.* 68, 42–49. doi: 10.1016/j.indcrop.2014.08.012

Do, H. D. K., Kim, C., Chase, M. W., and Kim, J. H. (2020). Implications of plastome evolution in the true lilies (monocot order Liliales). *Mol. Phylogenet. Evol.* 148:106818. doi: 10.1016/j.ympev.2020.106818

Dong, W., Liu, Y., Xu, C., Gao, Y., Yuan, Q., Suo, Z., et al. (2021). Chloroplast phylogenomic insights into the evolution of *Distylium* (Hamamelidaceae). *BMC Genom.* 22:293. doi: 10.1186/s12864-021-07590-6

Dong, W., Xu, C., Li, C., Sun, J., Zuo, Y., Shi, S., et al. (2015). ycf1, the most promising plastid DNA barcode of land plants. *Sci. Rep.* 5, 8348. doi: 10.1038/srep08348

Downie, S. R., and Jansen, R. K. (2015). A comparative analysis of whole plastid genomes from the Apiales: expansion and contraction of the inverted repeat, mitochondrial to plastid transfer of DNA, and identification of highly divergent noncoding regions. *Syst. Bot.* 40, 336–351. doi: 10.1600/036364415x686620

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.

Dugas, D. V., Hernandez, D., Koenen, E. J. M., Schwarz, E., Straub, S., Hughes, C. E., et al. (2015). Mimosoid legume plastome evolution: IR expansion, tandem repeat expansions, and accelerated rate of evolution in clpP. *Sci. Rep.* 5:16958. doi: 10.1038/srep16958

Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., and Dubchak, I. (2004). VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* 32, W273–W279. doi: 10.1093/nar/gkh458

Friis, I. (1989). "The Urticaceae: a systematic review," in *Evolution, Systematics, and Fossil History of the Hamamelidae*, eds P. R. Crane and S. Blackmore (Oxford: Clarendon Press), 285–308.

Friis, I. (1993). "Urticaceae," in *The Families and Genera of Vascular Plants*, eds K. Kubitzki, J. G. Rohwer, and V. Bittrich (Berlin: Springer-Verlag), 612–630. doi: 10.1007/978-3-662-02899-5_76

Fu, L. F., Xin, Z. B., Wen, F., Li, S., and Wei, Y. G. (2019). Complete chloroplast genome sequence of *Elatostema dissectum* (Urticaceae). *Mitochondrial DNA B* 4, 838–839. doi: 10.1080/23802359.2019.1567292

Gregory, T. R. (2005). DNA barcoding does not compete with taxonomy. *Nature* 434:1067. doi: 10.1038/4341067b

Grosse-Veldmann, B., Nürk, N. M., Smissen, R., Breitwieser, I., Quandt, D., and Weigend, M. (2016). Pulling the sting out of nettle systematics – a comprehensive phylogeny of the genus *Urtica* L. (Urticaceae). *Mol. Phylogenet. Evol.* 102, 9–19. doi: 10.1016/j.ympev.2016.05.019

Gu, C., Tembrock, L. R., Johnson, N. G., Simmons, M. P., and Wu, Z. (2016). The complete plastid genome of *Lagerstroemia fauriei* and loss of rpl2 intron from *Lagerstroemia* (Lythraceae). *PLoS One* 11:e0150752. doi: 10.1371/journal.pone.0150752

Guisinger, M. M., Chumley, T. W., Kuehl, J. V., Boore, J. L., and Jansen, R. K. (2010). Implications of the plastid genome sequence of typha (typhaceae, poales) for understanding genome evolution in poaceae. *J. Mol. Evol.* 70, 149–166. doi: 10.1007/s00239-009-9317-3

Gurung, A., Flanigan, H., Ghimeray, A., Bista, R., and Gunrung, O. (2012). Traditional knowledge of processing and use of the himalayan giant nettle (*Girardinia diversifolia* (Link) Friis) among the gurungs of sikles, Nepal. *Ethnobot. Res. App.* 10, 167–174. doi: 10.1234/era.v10i0.622

Haberle, R. C., Fourcade, H. M., Boore, J. L., and Jansen, R. K. (2008). Extensive rearrangements in the chloroplast genome of *Trachelium caeruleum* are associated with repeats and tRNA genes. *J. Mol. Evol.* 66, 350–361. doi: 10.1007/s00239-008-9086-4

Hadiah, J. T., Conn, B. J., and Quinn, C. J. (2008). Infra-familial phylogeny of Urticaceae, using chloroplast sequence data. *Aust. Syst. Bot.* 21, 375–385. doi: 10.1071/Sb08041

Huang, H., Shi, C., Liu, Y., Mao, S. Y., and Gao, L. Z. (2014). Thirteen *Camellia* chloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships. *BMC Evol. Biol.* 14:151. doi: 10.1186/1471-2148-14-151

Huang, X., Deng, T., Moore, M. J., Wang, H., Li, Z., Lin, N., et al. (2019). Tropical asian origin, boreotropical migration and long-distance dispersal in nettles (Urticeae, Urticaceae). *Mol. Phylogenet. Evol.* 137, 190–199. doi: 10.1016/j.ympev.2019.05.007

Huelsenbeck, J. P. (1995). Performance of phylogenetic methods in simulation. *Syst. Biol.* 44, 17–48. doi: 10.2307/2413481

Jansen, R. K., Raubeson, L. A., Boore, J. L., dePamphilis, C. W., Chumley, T. W., Haberle, R. C., et al. (2005). Methods for obtaining and analyzing whole chloroplast genome sequences. *Meth. Enzymol.* 395, 348–384. doi: 10.1016/S0076-6879(05)95020-9

Jansen, R. K., and Ruhlman, T. A. (2012). "Plastid genomes of seed plants," in *Genomics of Chloroplasts and Mitochondria*, eds R. Bock and V. Knoop (Dordrecht: Springer), 103–126. doi: 10.1007/978-94-007-2920-9_5

Jansen, R. K., Wojciechowski, M. F., Sanniyasi, E., Lee, S. B., and Daniell, H. (2008). Complete plastid genome sequence of the chickpea (*Cicer arietinum*) and the phylogenetic distribution of rps12 and clpP intron losses among legumes (Leguminosae). *Mol. Phylogenet. Evol.* 48, 1204–1217. doi: 10.1016/j.ympev.2008.06.013

Jin, X. F., Zhang, J., Lu, Y. F., Yang, W. W., and Chen, W. J. (2019). *Nanocnide zhejiangensis* sp. nov. (Urticaceae: Urticeae) from Zhejiang Province, East China. *Nord. J. Bot.* 37:e02339. doi: 10.1111/njb.02339

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

Kim, C., Deng, T., Chase, M., Zhang, D. G., Nie, Z. L., and Sun, H. (2015). Generic phylogeny and character evolution in Urticeae (Urticaceae) inferred from nuclear and plastid DNA regions. *Taxon* 64, 65–78. doi: 10.12705/641.20

Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J., and Giegerich, R. (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* 29, 4633–4642. doi: 10.1093/nar/29.22.4633

Li, J., Tang, J., Zeng, S., Han, F., Yuan, J., and Yu, J. (2021). Comparative plastid genomics of four *Pilea* (Urticaceae) species: insight into interspecific plastid genome diversity in *Pilea*. *BMC Plant Biol.* 21:25. doi: 10.1186/s12870-020-02793-7

Li, Y., Dong, Y., Liu, Y., Yu, X., Yang, M., and Huang, Y. (2020). Comparative analyses of *Euonymus* chloroplast genomes: genetic structure, screening for loci with suitable polymorphism, positive selection genes, and phylogenetic relationships within Celastrineae. *Front. Plant Sci.* 11:593984. doi: 10.3389/fpls.2020.593984

Lin, H. Y., Hao, Y. J., Li, J. H., Fu, C. X., Soltis, P. S., Soltis, D. E., et al. (2019). Phylogenomic conflict resulting from ancient introgression following species diversification in *Stewartia* s.l. (Theaceae). *Mol. Phylogenet. Evol.* 135, 1–11. doi: 10.1016/j.ympev.2019.02.018

Liu, D. K., Tu, X. D., Zhao, Z., Zeng, M. Y., Zhang, S., Ma, L., et al. (2020a). Plastid phylogenomic data yield new and robust insights into the phylogeny of *Cleisostoma–Gastrochilus* clades (Orchidaceae, Aeridinae). *Mol. Phylogenet. Evol.* 145:106729. doi: 10.1016/j.ympev.2019.106729

Liu, Q., Li, X., Li, M., Xu, W., Schwarzacher, T., and Heslop-Harrison, J. S. (2020b). Comparative chloroplast genome analyses of *Avena*: insights into evolutionary dynamics and phylogeny. *BMC Plant Biol.* 20:406. doi: 10.1186/s12870-020-02621-y

Liu, X., Chang, E. M., Liu, J. F., Huang, Y. N., Wang, Y., Yao, N., et al. (2019). Complete chloroplast genome sequence and phylogenetic analysis of *Quercus bawanglingensis* Huang, Li et Xing, a vulnerable oak tree in China. *Forests* 10:587. doi: 10.3390/f10070587

Luo, X., Li, L. L., Zhang, S. S., Lu, J. L., Zeng, Y., Zhang, H. Y., et al. (2011). Therapeutic effects of total coumarins from *Urtica dentata* hand on collagen-induced arthritis in Balb/c mice. *J. Ethnopharmacol.* 138, 523–529. doi: 10.1016/j.jep.2011.09.050

Ma, J., Yang, B., Zhu, W., Sun, L., Tian, J., and Wang, X. (2013). The complete chloroplast genome sequence of *Mahonia bealei* (Berberidaceae) reveals a significant expansion of the inverted repeat and phylogenetic relationship with other angiosperms. *Gene* 528, 120–131. doi: 10.1016/j.gene.2013.07.037

Mahlangeni, N. T., Moodley, R., and Jonnalagadda, S. B. (2020). Nutritional value, antioxidant and antidiabetic properties of nettles (*Laportea alatipes* and *Obetia tenax*). *Sci. Rep.* 10:9762. doi: 10.1038/s41598-020-67055-w

McDonald, M. J., Wang, W. C., Huang, H. D., and Leu, J. Y. (2011). Clusters of nucleotide substitutions and insertion/deletion mutations are associated with repeat sequences. *PLoS Biol.* 9:e1000622. doi: 10.1371/journal.pbio.1000622

Momo, C. E., Oben, J. E., Tazoo, D., and Dongo, E. (2006). Antidiabetic and hypolipidaemic effects of a methanol/methylene-chloride extract of *Laportea ovalifolia* (Urticaceae), measured in rats with alloxan-induced diabetes. *Ann. Trop. Med. Parasitol.* 100, 69–74. doi: 10.1179/136485906X78517

Naciri, Y., and Linder, H. P. (2015). Species delimitation and relationships: the dance of the seven veils. *Taxon* 64, 3–16. doi: 10.12705/641.24

Nicola, M. V., Johnson, L. A., and Pozner, R. (2019). Unraveling patterns and processes of diversification in the South Andean-Patagonian *Nassauvia* subgenus *Strongyloma* (Asteraceae, Nassauvieae). *Mol. Phylogenet. Evol.* 136, 164–182. doi: 10.1016/j.ympev.2019.03.004

Oyebanji, O., Zhang, R., Chen, S. Y., and Yi, T. S. (2020). New insights into the plastome evolution of the Millettioid/Phaseoloid clade (Papilionoideae, Leguminosae). *Front. Plant Sci.* 11:151–151. doi: 10.3389/fpls.2020.00151

Qu, X. J., Moore, M. J., Li, D. Z., and Yi, T. S. (2019). PGA: a software package for rapid, accurate, and flexible batch annotation of plastomes. *Plant Methods* 15:50. doi: 10.1186/s13007-019-0435-7

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

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., Hohna, S., et al. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542. doi: 10.1093/sysbio/sys029

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

Schattner, P., Brooks, A. N., and Lowe, T. M. (2005). The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.* 33, W686–W689. doi: 10.1093/nar/gki366

Schwarz, E. N., Ruhlman, T. A., Sabir, J. S. M., Hajrah, N. H., Alharbi, N. S., Al-Malki, A. L., et al. (2015). Plastid genome sequences of legumes reveal parallel inversions and multiple losses of rps16 in papilionoids. *J. Syst. Evol.* 53, 458–468. doi: 10.1111/jse.12179

Sharan Shrestha, S., Sut, S., Ferrarese, I., Barbon Di Marco, S., Zengin, G., De Franco, M., et al. (2020). Himalayan nettle *Girardinia diversifolia* as

a candidate ingredient for pharmaceutical and nutraceutical applications-phytochemical analysis and in vitro bioassays. *Molecules* 25:1563. doi: 10.3390/molecules25071563

Silverio, R. M. A. V., Vieira, L. D. N., Antonio de Baura, V., Balsanelli, E., Maltempi, de Souza, E., et al. (2021). Plastid phylogenomics of Pleurothallidinae (Orchidaceae): conservative plastomes, new variable markers, and comparative analyses of plastid, nuclear, and mitochondrial data. *PLoS One* 16:e0256126. doi: 10.1371/journal.pone.0256126

Simmonds, S. E., Smith, J. F., Davidson, C., and Buerki, S. (2021). Phylogenetics and comparative plastome genomics of two of the largest genera of angiosperms, Piper and *Peperomia* (Piperaceae). *Mol. Phylogenet. Evol.* 163:107229. doi: 10.1016/j.ympev.2021.107229

Singh, S. C., and Shrestha, R. (1988). *Girardinia diversifolia* (Urticaceae), a non-conventional fiber resource in Nepal. *Econ. Bot.* 42, 445–447.

Sloan, D. B., Triant, D. A., Forrester, N. J., Bergner, L. M., Wu, M., and Taylor, D. R. (2014). A recurring syndrome of accelerated plastid genome evolution in the angiosperm tribe *Sileneae* (Caryophyllaceae). *Mol. Phylogenet. Evol.* 72, 82–89. doi: 10.1016/j.ympev.2013.12.004

Som, A. (2014). Causes, consequences and solutions of phylogenetic incongruence. *Brief. Bioinform.* 16, 536–548. doi: 10.1093/bib/bbu015

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

Stevens, P. F. (2017). *Angiosperm Phylogeny Website. Version 14, July 2017*. Available online at: http://www.mobot.org/MOBOT/research/APweb (accessed November 9, 2021).

Sullivan, J., and Joyce, P. (2005). Model selection in phylogenetics. *Annu. Rev. Ecol. Evol. Syst.* 36, 445–466. doi: 10.1146/annurev.ecolsys.36.102003.152633

Sytsma, K. J., Morawetz, J., Pires, J. C., Nepokroeff, M., Conti, E., Zjhra, M., et al. (2002). Urticalean rosids: circumscription, rosid ancestry, and phylogenetics based on *rbcL*, *trnL*-F, and *ndhF* sequences. *Am. J. Bot.* 89, 1531–1546. doi: 10.3732/ajb.89.9.1531

Tanti, B., Buragohain, A., Gurung, L., Kakati, D., Das, A. K., and Borah, S. P. (2010). Assessment of antimicrobial and antioxidant activities of *Dendrocnide sinuata* (Blume) Chew leaves-a medicinal plant used by ethnic communities of North East India. *IJNPR* 1, 17–21.

Thiel, T., Michalek, W., Varshney, R. K., and Graner, A. (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor. Appl. Genet.* 106, 411–422. doi: 10.1007/s00122-002-1031-0

Wang, R. N., Milne, R. I., Du, X. Y., Liu, J., and Wu, Z. Y. (2020b). Characteristics and mutational hotspots of plastomes in *Debregeasia* (Urticaceae). *Front. Genet.* 11:729. doi: 10.3389/fgene.2020.00729

Wang, H. X., Liu, H., Moore, M. J., Landrein, S., Liu, B., Zhu, Z. X., et al. (2020a). Plastid phylogenomic insights into the evolution of the Caprifoliaceae s.l. (Dipsacales). *Mol. Phylogenet. Evol.* 142:106641. doi: 10.1016/j.ympev.2019.106641

Wang, J. H., Moore, M. J., Wang, H., Zhu, Z. X., and Wang, H. F. (2021). Plastome evolution and phylogenetic relationships among Malvaceae subfamilies. *Gene* 765:145103. doi: 10.1016/j.gene.2020.145103

Wang, W. T., and Chen, C. J. (1995). *Flora Reipublicae Popularis Sinicae*, Vol. 23. Beijing: Science Press.

Wells, T., Maurin, O., Dodsworth, S., Friis, I., Cowan, R., Epitawalage, N., et al. (2021). Combination of Sanger and target-enrichment markers supports revised generic delimitation in the problematic '*Urera* clade' of the nettle family (Urticaceae). *Mol. Phylogenet. Evol.* 158:107008. doi: 10.1016/j.ympev.2020.107008

Wendel, J. F., and Doyle, J. J. (1998). "Phylogenetic incongruence: window into genome history and molecular evolution," in *Molecular Systematics of Plants II*, eds P. S. Soltis, D. E. Soltis, and J. J. Doyle (Boston: Kluwer Academic Publishing), 265–296. doi: 10.1007/978-1-4615-5419-6_10

Wick, R. R., Schultz, M. B., Zobel, J., and Holt, K. E. (2015). Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* 31, 3350–3352. doi: 10.1093/bioinformatics/btv383

Wicke, S., Schneeweiss, G. M., dePamphilis, C. W., Muller, K. F., and Quandt, D. (2011). The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Mol. Biol.* 76, 273–297. doi: 10.1007/s11103-011-9762-4

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

Wu, Z. Y., Liu, J., Provan, J., Wang, H., Chen, C. J., Cadotte, M. W., et al. (2018). Testing Darwin's transoceanic dispersal hypothesis for the inland nettle family (Urticaceae). *Ecol. Lett.* 21, 1515–1529. doi: 10.1111/ele.13132

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

Xue, J., Wang, S., and Zhou, S. L. (2012). Polymorphic chloroplast microsatellite loci in *Nelumbo* (Nelumbonaceae). *Am. J. Bot.* 99, e240–e244. doi: 10.3732/ajb.1100547

Xue, S., Shi, T., Luo, W., Ni, X., Iqbal, S., Ni, Z., et al. (2019). Comparative analysis of the complete chloroplast genome among *Prunus mume*, P. *armeniaca*, and P. *salicina*. *Hortic. Res.* 6:89. doi: 10.1038/s41438-019-0171-1

Yan, M., Fritsch, P. W., Moore, M. J., Feng, T., Meng, A., Yang, J., et al. (2018). Plastid phylogenomics resolves infrafamilial relationships of the Styracaceae and sheds light on the backbone relationships of the Ericales. *Mol. Phylogenet. Evol.* 121, 198–211. doi: 10.1016/j.ympev.2018.01.004

Yang, Y., Zhou, T., Duan, D., Yang, J., Feng, L., and Zhao, G. (2016). Comparative analysis of the complete chloroplast genomes of five *Quercus* species. *Front. Plant Sci.* 7:959. doi: 10.3389/fpls.2016.00959

Yao, G., Jin, J. J., Li, H. T., Yang, J. B., Mandala, V. S., Croley, M., et al. (2019). Plastid phylogenomic insights into the evolution of Caryophyllales. *Mol. Phylogenet. Evol.* 134, 74–86. doi: 10.1016/j.ympev.2018.12.023

Zhang, X. F., Landis, J. B., Wang, H. X., Zhu, Z. X., and Wang, H. F. (2021). Comparative analysis of chloroplast genome structure and molecular dating in Myrtales. *BMC Plant Biol.* 21:219. doi: 10.1186/s12870-021-02985-9

Zheng, S., Poczai, P., Hyvonen, J., Tang, J., and Amiryousefi, A. (2020). Chloroplot: an online program for the versatile plotting of organelle genomes. *Front. Genet.* 11:576124. doi: 10.3389/fgene.2020.576124

Zhu, A., Guo, W., Gupta, S., Fan, W., and Mower, J. P. (2016). Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. *New Phytol.* 209, 1747–1756. doi: 10.1111/nph.13743

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

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

Copyright © 2022 Ogoma, Liu, Stull, Wambulwa, Oyebanji, Milne, Monro, Zhao, Li and Wu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 23 February 2022
doi: 10.3389/fpls.2022.823190

Highly Resolved Papilionoid Legume Phylogeny Based on Plastid Phylogenomics

In-Su Choi[1], Domingos Cardoso[2]*, Luciano P. de Queiroz[3], Haroldo C. de Lima[4], Chaehee Lee[5], Tracey A. Ruhlman[5], Robert K. Jansen[5,6] and Martin F. Wojciechowski[1]*

[1] School of Life Sciences, Arizona State University, Tempe, AZ, United States, [2] National Institute of Science and Technology in Interdisciplinary and Transdisciplinary Studies in Ecology and Evolution (INCT IN-TREE), Instituto de Biologia, Universidade Federal da Bahia, Salvador, Brazil, [3] Department of Biological Sciences, Universidade Estadual de Feira de Santana, Feira de Santana, Brazil, [4] Instituto de Pesquisas Jardim Botânico do Rio de Janeiro, Rio de Janeiro, Brazil, [5] Department of Integrative Biology, University of Texas at Austin, Austin, TX, United States, [6] Center of Excellence for Bionanoscience Research, King Abdulaziz University (KAU), Jeddah, Saudi Arabia

OPEN ACCESS

Edited by:
Gerald Matthias Schneeweiss, University of Vienna, Austria

Reviewed by:
Ashley N. Egan, Utah Valley University, United States
Ricarda Riina, Real Jardín Botánico, Spanish National Research Council (CSIC), Spain

***Correspondence:**
In-Su Choi
86ischoi@gmail.com
Domingos Cardoso
cardosobot@gmail.com

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

Received: *26 November 2021*
Accepted: *31 January 2022*
Published: *23 February 2022*

Citation:
Choi I-S, Cardoso D, de Queiroz LP, de Lima HC, Lee C, Ruhlman TA, Jansen RK and Wojciechowski MF (2022) Highly Resolved Papilionoid Legume Phylogeny Based on Plastid Phylogenomics. Front. Plant Sci. 13:823190. doi: 10.3389/fpls.2022.823190

Comprising 501 genera and around 14,000 species, Papilionoideae is not only the largest subfamily of Fabaceae (Leguminosae; legumes), but also one of the most extraordinarily diverse clades among angiosperms. Papilionoids are a major source of food and forage, are ecologically successful in all major biomes, and display dramatic variation in both floral architecture and plastid genome (plastome) structure. Plastid DNA-based phylogenetic analyses have greatly improved our understanding of relationships among the major groups of Papilionoideae, yet the backbone of the subfamily phylogeny remains unresolved. In this study, we sequenced and assembled 39 new plastomes that are covering key genera representing the morphological diversity in the subfamily. From 244 total taxa, we produced eight datasets for maximum likelihood (ML) analyses based on entire plastomes and/or concatenated sequences of 77 protein-coding sequences (CDS) and two datasets for multispecies coalescent (MSC) analyses based on individual gene trees. We additionally produced a combined nucleotide dataset comprising CDS plus *matK* gene sequences only, in which most papilionoid genera were sampled. A ML tree based on the entire plastome maximally supported all of the deep and most recent divergences of papilionoids (223 out of 236 nodes). The Swartzieae, ADA (Angylocalyceae, Dipterygeae, and Amburaneae), Cladrastis, Andira, and Exostyleae clades formed a grade to the remainder of the Papilionoideae, concordant with nine ML and two MSC trees. Phylogenetic relationships among the remaining five papilionoid lineages (Vataireoid, *Dermatophyllum*, Genistoid s.l., Dalbergioid s.l., and Baphieae + Non-Protein Amino Acid Accumulating or NPAAA clade) remained uncertain, because of insufficient support and/or conflicting relationships among trees. Our study fully resolved most of the deep nodes of Papilionoideae, however, some relationships require further exploration. More genome-scale data and rigorous analyses are needed to disentangle phylogenetic relationships among the five remaining lineages.

Keywords: deep evolution, Meso-Papilionoideae, plastid genome, Papilionoideae, Leguminosae

INTRODUCTION

Advances in next-generation sequencing and computational resources have enabled unparalleled phylogenomic analyses. These studies have deepened our understanding of evolutionary relationships across many branches in the plant Tree of Life, from the most recalcitrant deep relationships at and within the family level (e.g., Xi et al., 2012; Goremykin et al., 2015; Duvall et al., 2020; Koenen et al., 2020a; Yang et al., 2020; Antonelli et al., 2021; Orton et al., 2021; Schneider et al., 2021; Serna-Sánchez et al., 2021) to long, unresolved radiations at the species level (e.g., Nicholls et al., 2015; Welch et al., 2016; Villaverde et al., 2018; Thode et al., 2020; Pereira et al., 2021). While massive amounts of plastid genome (plastome) sequence data have filled the family level sampling gap for angiosperms (e.g., Li et al., 2021), infra-family levels remain less well covered. This is particularly true of the economically important, ecologically successful, morphologically diverse, species-rich legume family Fabaceae (Leguminosae), from which the plastomes of only 319 species in 184 genera have been deposited in the GenBank database[1] (Accessed Sep. 09, 2021) thus far, of the more than 22,000 species and 770 genera in six subfamilies (LPWG, 2017, 2021).

Fabaceae is one of the most spectacular examples of diversification among flowering plants. Many legumes are not only ecologically dominant across major tropical and subtropical biomes (Schrire et al., 2005; DRYFLOR, 2016; LPWG, 2017) but symbiotically associated with nitrogen-fixing bacteria *via* root nodules (Sprent et al., 2017), amplifying the family's importance for food security, sustainable agriculture, and ecosystem function (Food and Agriculture Organization[2]; LPWG, 2013; Yahara et al., 2013). The successful radiation of legumes is thought to be associated with plant defense strategies against herbivores, diverse, intimate ecological interactions involving ant-housing domatia and ant-feeding extrafloral nectaries (Janzen, 1966; Marazzi and Sanderson, 2010; Chomicki et al., 2015; Marazzi et al., 2019), an extraordinary range of floral forms (Tucker, 2003a) and pollination mechanisms (Arroyo, 1981). The family provides a wide diversity of secondary metabolites (alkaloids, flavonoids, lignans, tannins, terpenoids, benzofuranoids, and non-proteinogenic amino acids such as canavanine; Bisby et al., 1994; Kursar et al., 2009; Wink, 2013). The family is also an excellent model to reveal the patterns and processes of plastome structural evolution, since its taxa have undergone several dramatic rearrangements involving inversions of large blocks of sequence, contraction, loss, and regain of the inverted repeat (IR), gene/intron loss and repeat accumulation (e.g., Palmer et al., 1987; Lavin et al., 1990; Doyle et al., 1996; Jansen et al., 2008; Martin et al., 2014; Schwarz et al., 2015; Choi and Choi, 2017; Choi et al., 2019; Zhang et al., 2020; Charboneau et al., 2021; Lee et al., 2021). One of the most striking examples is a 50-kb inversion situated in the plastome large single-copy region (LSC) that is shared by the vast majority of subfamily Papilionoideae (papilionoids) (Doyle et al., 1996; Pennington et al., 2001; Wojciechowski et al., 2004; Cardoso et al., 2012a; LPWG, 2013). The 50 kb-inversion was long considered an unequivocal molecular synapomorphy for this clade, but recently at least three species of *Sesbania* Adans. were shown to have completely reverted the 50-kb sequence, resulting in essentially the same gene order as found in the earliest-diverging papilionoids (Lee et al., 2021).

In addition to providing a model for plastome structural rearrangements, Papilionoideae, the largest legume subfamily with an estimated 501 genera and 14,000 species (LPWG, 2021), also exhibits an impressive morphological diversity (Lewis et al., 2005; LPWG, 2017). For example, the early diversification of the Papilionoideae is marked by multiple evolutionary shifts in floral architecture (Ireland et al., 2000; Pennington et al., 2001; Cardoso et al., 2012a, 2013a,b; Klitgård et al., 2013; Ramos et al., 2016; Castellanos et al., 2017). Genera that were traditionally classified in the "primitive" tribes Sophoreae and Swartzieae (e.g., Cowan, 1981; Polhill, 1981a) are now phylogenetically scattered among the early-branching lineages of Papilionoideae. Their flowers are morphologically variable, from actinomorphic (radial or polysymmetric) with five undifferentiated petals to zygomorphic (bilateral or monosymmetric) with the petals poorly differentiated, but also absent or restricted to just the adaxial standard petal, and with free, often numerous stamens (Pennington et al., 2000; Cardoso et al., 2012a). The main phylogenetic outcome from early, single molecular locus Papilionoideae phylogenies (Doyle et al., 1997; Pennington et al., 2001; Wojciechowski et al., 2004) was that the papilionate-flowered ancestors experienced high evolutionary lability during the initial diversification of the subfamily, refuting the notion that non-papilionate flowers represented signatures of antiquity (e.g., Arroyo, 1981; Polhill, 1981a; Tucker and Douglas, 1994).

The extraordinary evolutionary and ecological success of legumes during their more than 60 million years of diversification history (Lavin et al., 2005; Bruneau et al., 2008; Koenen et al., 2021) may be related to the macroevolutionary stability of the highly specialized papilionate flower (Cardoso et al., 2013a), beneficial associations with nodulating symbiotic bacteria (Sprent et al., 2017), ant feeding *via* extrafloral nectaries (Marazzi et al., 2012, 2019) and/or secondary metabolite accumulation (Wink, 2013). Determining the emergence and influence of these phenomena in the evolutionary history of legumes requires a well-sampled and fully resolved phylogeny.

Previous phylogenetic studies using the coding sequences of the plastid *rbcL* (e.g., Doyle et al., 1997; Kajita et al., 2001) and *matK* genes (e.g., Wojciechowski et al., 2004; Cardoso et al., 2012a, 2013a, 2015; Ramos et al., 2016; Castellanos et al., 2017; LPWG, 2017; Queiroz et al., 2017), as well as a supermatrix approach (McMahon and Sanderson, 2006), sampled densely across the Papilionoideae, revealed many new clades, unexpected generic re-alignments, and placed several taxonomically orphan genera. However, these studies left the Papilionoideae backbone phylogeny and the placement of several evolutionary key genera largely unresolved. Such is the case with the small temperate North American genus *Dermatophyllum* Scheele. Apart from the Genistoid s.l. clade, *Dermatophyllum* is the only lineage of legumes known to accumulate a variety of quinolizidine

[1] https://www.ncbi.nlm.nih.gov/genbank/

[2] http://www.fao.org/pulses-2016/en/

alkaloids (QA), a class of alkaloids mainly distributed in these two papilionoid lineages with some phylogenetically scattered occurrences within other angiosperm families (Bisby et al., 1994; Kite and Pennington, 2003; Lee et al., 2013; Wink, 2013). Determining the sister relationship of this enigmatic, isolated genus with respect to the genistoids is fundamental to answer whether the evolution of such an important secondary metabolite in papilionoids had a single or multiple origins. Also unresolved is the polytomy within the large 50-kb-inversion clade involving the species-rich Dalbergioid s.l. and Non-Protein Amino Acid Accumulating (NPAAA) clades, as well as the florally heterogeneous Andira, Exostyleae, and Vataireoid clades. Previous phylogenies with broad taxon sampling resolved the earliest divergences of the Papilionoideae involving the swartzioids and the ADA clade (Angylocalyceae, Dipterygeae, and Amburaneae), albeit with low support values. These clades are often interchangeably shown as sister to the remainder of the subfamily (Cardoso et al., 2012a, 2013a, 2015; Ramos et al., 2016; Zhang et al., 2020; Zhao et al., 2021).

Recent advances in deep- and genus-level legume phylogenomics using both plastid and nuclear genes (Cannon et al., 2015; LPWG, 2017; Vatanparast et al., 2018; Ojeda et al., 2019; Koenen et al., 2020a,b, 2021; Oyebanji et al., 2020; Zhang et al., 2020; Zhao et al., 2021) have contributed to the resolution of some of the obscure deep relationships in the Papilionoideae. Still, many morphological key genera in the early-diverging lineages outside the agriculturally important NPAAA clade (Cardoso et al., 2012a, 2013a) have not been evaluated in any phylogenomic study. Among the legume phylogenomic analyses that more thoroughly investigated the Papilionoideae (Koenen et al., 2020a; Zhang et al., 2020; Zhao et al., 2021), only Zhao et al. (2021) have greatly improved taxon sampling within the early-diverging clades outside the NPAAA clade. In that study, more than 1,500 nuclear genes from transcriptome and genome assemblies of 217 papilionoid genera were explored. Previous plastome-inferred legume phylogenies have sampled just 32 (Koenen et al., 2020a) and 48 (Zhang et al., 2020) papilionoid genera. These studies all left important gaps, for example, regarding the placement of *Dermatophyllum*, either because no representative species were sampled or because statistical support for its sister relationship was relatively low. Thus, we still lack a clearer, more focused picture of Papilionoideae evolutionary history from comprehensively sampled plastome data.

In this study, we used a denser taxon sampling of plastome sequence data with a focus on the early-branching Papilionoideae. Despite the fact that plastome data may produce conflicting topologies (e.g., Gonçalves et al., 2019; Walker et al., 2019; Koenen et al., 2020a), examining them thoroughly can help to clarify the phylogenetic signal in concatenated analyses of a large number of plastid loci (Gonçalves et al., 2019; Zhang et al., 2020). For example, the inclusion of protein-coding genes only, removal of ambiguously aligned regions, or the use of coalescent methods (e.g., Gonçalves et al., 2019) may be helpful to arrive at phylogenetically accurate topologies. Here, we aimed to explore phylogenetic signals across diverse plastome-derived datasets together with cross-genomic comparisons of mitochondrial and nuclear data and shed light on key evolutionary relationships related to a number of questions that have persisted in the subfamily.

MATERIALS AND METHODS

Plant Material Sampling

Our taxon sampling strategy was designed to cover the major papilionoid clades recognized by Cardoso et al. (2012a, 2013a), especially those in need of additional investigation. Thirty-eight taxa were collected from Brazil and the United States of America. Seeds of *Phaseolus acutifolius* A. Gray were acquired from the Desert Legume Program (University of Arizona, Tucson) and grown in the University of Texas at Austin (UT-Austin) greenhouse. Leaf samples were collected, flash-frozen in liquid nitrogen and stored at −80°C before DNA isolation except the silica-gel dried leaf tissues of *Astragalus canadensis* var. *brevidens* (Gand.) Barneby. Voucher specimens are housed at the Billie L. Turner Plant Resource Center at UT-Austin (TEX-LL), the Arizona State University (ASU), the Universidade Estadual de Feira de Santana (HUEFS), and the Jardim Botânico do Rio de Janeiro (RB) herbaria. Collection information for a total of 39 taxa is available in **Supplementary Table 1**.

Next-Generation Sequencing and Plastid Genome Completion

Total genomic DNAs were extracted by the hexadecyltrimethylammonium bromide protocol, described in Doyle and Doyle (1987), or using the NucleoSpin Plant II, Mini Kit for DNA from plants (Macherey-Nagel, Düren, Germany). Next-generation sequencing (NGS) reads (2 × 150 bp) from ca. 300 bp insert libraries for 38 taxa were generated by the Beijing Genomics Institute (BGI; Shenzhen, China) using the MGI DNBseq^{TM} platform (MGI Tech Co., Shenzhen, China). Total genomic DNA of *A. canadensis* var. *brevidens* was sequenced using the Illumina NextSeq 500 at the Arizona State University Genomics Facility (2 × 151 bp; ca. 300 bp insert size). Plastomes were assembled and annotated by the methods described in Choi et al. (2019).

Plastid Phylogenomics

In addition to the complete plastomes generated in this study, we obtained the plastome sequence from at least a single species of each papilionoid genus currently available in GenBank (see text footnote 1) (Accessed May 24, 2021) or Dryad[3] (Accessed Aug. 24, 2020). In total, 244 taxa, with 237 papilionoids (174 genera) and seven non-papilionoids were collected as ingroup and outgroup, respectively (**Supplementary Table 2**). We produced nine sequence alignments as data sets for maximum likelihood (ML) analysis and two gene tree sets for species tree estimation based on the multispecies coalescent (MSC) approach (**Table 1** and **Supplementary Table 3**).

Datasets 1—3 (WP, WP_nogap, and WP_gb; see **Table 1**) were prepared based on whole plastome alignments. For the legume taxa with two copies of the large IR, one copy was deleted. To

[3]https://datadryad.org

align plastomes with different gene order, Mauve 2.3.1 (Darling et al., 2010) was used to detect locally colinear blocks (LCBs) relative to *Cercis canadensis* L. (KF856619). Sequence blocks of all taxa were rearranged to be colinear with *C. canadensis*. To avoid introducing non-orthologous sequences during the rearrangements, non-genic edges of LCBs were deleted. The intergenic regions that coincide with the end points of the 50-kb inversion and an adjacent gene encoding *rps16*, pseudogenized in many papilionoids (Schwarz et al., 2015; Choi and Choi, 2017; Lee et al., 2021), were deleted from all taxa. Complete plastomes were aligned by MAFFT v7.450 (Katoh and Standley, 2013) in Geneious Prime 2021.0.3[4] using default options. This raw alignment is designated as WP (dataset 1). The WP dataset was further refined by two different strategies. The WP_nogap (dataset 2) was prepared by deleting all indel regions using the "mask alignment" tool in Geneious Prime, as described in Orton et al. (2021). The WP_gb (dataset 3) was prepared from the WP dataset using Gblocks 0.91b (Castresana, 2000) using default options allowing selection of conserved sequence blocks without indel regions exclusively.

Datasets 4—8 (CD_nt, CD_nt_gb, CD_nt_dg, CD_aa, and CD_aa_gb; see **Table 1**) were prepared based on the protein-coding sequences (CDS) for 77 genes (**Supplementary Table 4**) extracted from each plastome. The 77 CDS included putatively pseudogenized genes with few mutations that could represent heteroplasmic variations (Park et al., 2020). In plastomes where the intactness of *rps16* was questionable, the pseudogene was unsampled. A single nucleotide (A, T, C, G, or N) was introduced or deleted to fit the reading frame when a gene included a premature stop codon due to indel polymorphism. Nucleotide sequences of each CDS dataset were aligned by MAFFT using the translation align option available in Geneious Prime. Finally, poorly aligned regions of each CDS alignment were manually adjusted or deleted. A concatenated aligned matrix of 77 CDS sequences was produced using the R package catGenes[5]. This concatenated nucleotide alignment (dataset 4) was designated CD_nt. This data set was further refined by Gblocks using default options resulting in CD_nt_gb (dataset 5). To avoid the problem of nucleotide compositional heterogeneity among taxa, the synonymous substitution sites in the CD_nt (dataset 4) were degenerated using Degen ver. 1.4.[6] (Regier et al., 2010) and the result was designated CD_nt_dg (dataset 6). Datasets 7 and 8 were prepared based on translated amino acid sequences (AA) of CD_nt (dataset 4): Dataset 7 (CD_aa) was aligned without trimming and dataset 8 (CD_aa_gb) was trimmed using Gblocks with default options.

To more broadly examine the phylogenetic relationships of papilionoid legumes we included taxa that were not sampled in our plastome datasets but were previously sequenced for *matK*. We produced dataset 9 (CD_matK_cb; see **Table 1**) by combining CD_nt (dataset 4) and a *matK*-only dataset of 534 nucleotide sequences. For this alignment the taxa lacking plastome-scale data were represented by the *matK* coding region while their 76 CDS were coded as missing data. In total, dataset 9 included 771 papilionoids (478 genera) (**Supplementary Table 5**). The dataset included *matK* sequences used in LPWG (2017) and excluded duplications and short sequence fragments (< 800 bp).

For each dataset (1—9), ML analysis was conducted using IQ-TREE 1.6.12 (Nguyen et al., 2015) with 1,000 bootstrap (BS) replications, and a best-fit nucleotide substitution model was automatically selected based on the Bayesian information criterion.

Datasets 10 and 11 (see **Table 1**) consisted of sets of individual ML trees based on each CDS for species tree estimation with MSC approach using ASTRAL-III (Zhang et al., 2018). The ML analyses for each CDS were conducted as described above. To decrease error rate, branches with BS value lower than 10 were contracted as suggested in Zhang et al. (2018). Dataset 10 included all 77 individual trees. The individual trees with strong phylogenetic signals (average BS value > 85) were selected and this set of trees was designated dataset 11. The local posterior probability (LPP) and quartet score (QS) were calculated for all nodes.

Visualization and editing of phylogenetic trees were conducted using Interactive Tree Of Life (iTOL; Letunic and Bork, 2021), ggtree package (Yu et al., 2017), and custom R scripts.

[4]https://www.geneious.com

TABLE 1 | Descriptions for 11 datasets used for plastid phylogenomic analyses in this study.

Tree number	Dataset name	Description
1	WP	Whole plastid genome alignment
2	WP_nogap	WP dataset with all indels removed
3	WP_gb	WP dataset with only conserved sequence blocks
4	CD_nt	Concatenation of 77 CDS nucleotide alignments
5	CD_nt_gb	CD_nt dataset with only conserved sequence blocks
6	CD_nt_dg	CD_nt dataset with degenerated synonymous substitutions sites
7	CD_aa	Concatenation of 77 CDS amino acid alignments
8	CD_aa_gb	CD_aa dataset with only conserved sequence blocks
9	CD_matK_cb	Combined dataset of CD_nt and *matK*-only dataset
10	77 gene trees	All 77 individual gene trees
11	26 gene trees	26 gene trees with a mean bootstrap value higher than 85

RESULTS

In total, we included plastomes from 244 species (**Supplementary Table 2**), 39 of which were sequenced for this study (**Supplementary Table 1**), and *matK* sequences from 534 species (**Supplementary Table 5**). This taxon sampling scheme

[5]https://github.com/domingoscardoso/catGenes

[6]http://www.phylotools.com/

includes representatives of all 22 main, early-diverging lineages (19 in plastome-only sampling, without *Amphimas* Pierre ex Harms, *Aldina* Endl., and *Clathrotropis macrocarpa* Ducke), as recognized by Cardoso et al. (2012a), and spans a diversity of taxa exhibiting radially symmetrical to bilaterally symmetrical floral architecture (**Figure 1**).

In total, 39 papilionoid plastomes were assembled and included in our analysis (**Supplementary Table 1**). Plastomes varied from 123,013 (*Astragalus canadensis* var. *brevidens*) to 168,148 bp (*Myrospermum sousanum* A.Delgado & M. C.Johnst.) in unit length. The read depth varied from 409 to 6,213×. Together with previously sequenced plastid data, eight plastome-only and one combined (*matK*-only + CD_nt) datasets were prepared for our ML analyses of papilionoid legumes (**Tables 1** and **Supplementary Table 3**). The nine datasets for ML analyses varied in the values for the alignment length, proportion of gaps/ambiguities, and number of parsimony-informative sites (PIS) (**Figure 2A**). The highest values of the alignment length and PIS were shown in the WP dataset with 313,335 bp (60.3% gaps/ambiguities) and 101,445 sites, respectively, values substantially higher than the rest of the datasets. The original concatenation of all CDS (CD_nt) showed higher values for the alignment length and number of PIS, but a trimmed version (CD_nt_gb) was similar to plastome alignments without indel regions (WP_nogap) or with only conserved sequence blocks (WP_gb). Both modification of CD_nt (CD_nt_dg) and the combination of CD_nt with the *matK*-only dataset (CD_matK_cb) resulted in changes in the number of the PIS and gaps/ambiguity (**Supplementary Table 3**).

The datasets based on translated AAs showed the lowest values for alignment length and number of PIS. Datasets 10 and 11 (77 gene trees and 26 gene trees) employed for the MSC approach were prepared based on ML analyses for each individual CDS alignment (**Supplementary Table 4**). Length of each alignment varied from 90 bp (*petN*) to 11,175 (*ycf1*) bp. The lowest and highest mean BS values were shown from *psbN* (45.7) and ycf1 (97.9) (**Figure 2B** and **Supplementary Table 4**). There were 26 trees with average BS > 85 (i.e., dataset 11).

Based on 11 datasets, 11 trees were inferred and compared (**Figure 2C**). Tree 1 (WP dataset) resolved the most papilionoid nodes (223 out of 236, 93.5%). Trees 2—5 resolved a similar number of nodes with maximal support (BS = 100) (215—217, 91.1—91.9%). Trees 6—8 showed the lowest number of resolved nodes with maximal support (196—207, 83.1—87.7%) among ML trees. The many additional papilionoid nodes in Tree 9 (CDS and *matK*-only) resolved 580 out of 770 nodes (75.3%) with maximal support. Multispecies coalescence trees (10 and 11) resolved a similar number of nodes [192 (81.4%) and 193 (81.8%), respectively] with maximal support (LPP = 1.0).

Topological concordance with regard to 10 main papilionoid lineages (Swartzieae, ADA, Cladrastis, Andira, Exostyleae, *Dermatophyllum*, Vataireoid, Genistoid s.l., Dalbergioid s.l., and Baphieae + NPAAA clades) was assessed across the 11 trees (**Figure 3**). Topologies showing the five lineages Swartzieae, ADA, Cladrastis, Andira, and Exostyleae clades as successive sister groups to a monophyletic group comprising the remaining five papilionoid lineages (*Dermatophyllum*, Vataireoid, Genistoid s.l., Dalbergioid s.l., and Baphieae + NPAAA clades) were consistently retrieved from all trees. The three nodes related to the five remaining lineages resolved differently. The most frequently recovered relationship subdivided the group into two clades of (Dalbergioid s.l., Baphieae + NPAAA) and (*Dermatophyllum*, (Vataireoid, Genistoid s.l.)). Support and alternative topologies for these five lineages are shown in **Figure 4**. While all 11 trees resolved relationships, only the nodes inferred in Tree 1 had 100% BS support. Among the seven ML trees (except Tree 1) based on plastome datasets, Trees 7 and 8 from translated AA alignments showed slightly better support for the interrelationships of the five lineages. The MSC-based Tree 10 had only low LPP and QS for the main topology of the five lineages. The selection of 26 genes with high mean BS values as the input dataset (dataset 11) showed slightly increased LPP and QS for the monophyly of the five remaining lineages but does not substantially affect resolution of relationships among these lineages (Tree 11). To show specific relationships within and among the 10 key papilionoid lineages, Tree 5 along with support values of Trees 1 and 11 are presented as the main tree (**Figure 5**). In total, the 50 papilionoid genera that need further attention in plastid phylogenomics are highlighted in **Figure 6** based on the most comprehensive taxon sampling (CD_matK_cb). These 50 genera are either still phylogenetically enigmatic, are morphological key genera, or represent monospecific lineages.

DISCUSSION

Plastid Phylogenomic Signals Across the Papilionoid Legumes

Maximum Likelihood analysis of the whole plastome alignment (**Figure 4**) resolved all deep relationships among major early-diverging clades of the Papilionoideae with maximal support values. This approach for phylogenetic analyses using the entire plastome as a single locus was rigorously tested in Poaceae (Duvall et al., 2020; Orton et al., 2021). Duvall et al. (2020) demonstrated how tree topology and support values could vary according to the level of gap removal from the whole plastome alignment, and recommended excluding any gapped positions from machine-generated alignments. Apart from concerns as to whether it is reasonable to recognize the plastome as a single locus (Gonçalves et al., 2020; Doyle, 2021), one of the main and practical criticisms about whole plastome alignment without further filtration is that most gaps are located in AT-rich and low complexity regions, which are prone to high-level length polymorphisms in simple sequence repeats and falsely aligned non-orthologous sequences (Duvall et al., 2020). The much longer whole plastome alignment of 244 legumes (313,335 bp) compared to the alignment lacking indels (64,099 bp) is not surprising when considering the dynamic gene, intron, intergenic, and repeat content across papilionoids (e.g., Milligan et al., 1989; Doyle et al., 1995; Bailey et al., 1997; Cai et al., 2008; Jansen et al., 2008; Sabir et al., 2014; Sveinsson and Cronk, 2014; Schwarz et al., 2015; Choi et al., 2019, 2020a; Jin et al., 2019; Oyebanji et al., 2020; Lee et al., 2021). The WP alignment extended well beyond the length of a typical plastome

FIGURE 1 | Broad variation in floral architecture across the papilionoid legumes (Papilionoideae). The selected taxa represent genera from the early-branching lineages **(A–M)** outside the NPAAA (non-protein amino acid accumulating) clade. **(A)** *Swartzia acutifolia* (Swartzieae); **(B)** *Castanospermum australe* [Angylocalyceae, ADA clade (Angylocalyceae, Dipterygeae, and Amburaneae)]; **(C)** *Dipteryx magnifica* (Dipterygeae, ADA clade); **(D)** *Myrocarpus fastigiatus* (Amburaneae, ADA clade); **(E)** *Pickeringia montana* (Cladrastis clade); **(F)** *Harpalyce brasiliana* (Brongniartieae, Genistoid s.l.); **(G)** *Aldina latifolia* (Andira clade); **(H)** *Exostyles venusta* (Exostyleae); **(I)** *Dermatophyllum secundiflorum* (unresolved); **(J)** *Dalea mollis* (Amorpheae); **(K)** *Centrolobium microchaete* (Dalbergieae); **(L)** *Leptolobium dasycarpum* (Leptolobieae, Genistoid s.l.); **(M)** *Lupinus sericeus* (Genisteae, Genistoid s.l.); **(N)** *Robinia neomexicana* (Robinieae, NPAAA); **(O)** *Astragalus mollissimus* (Astragalean, NPAAA); Photos by Domingos Cardoso **(A–D,F–H,K–M)** and Martin F. Wojciechowski **(E,I,J,N,O)**.

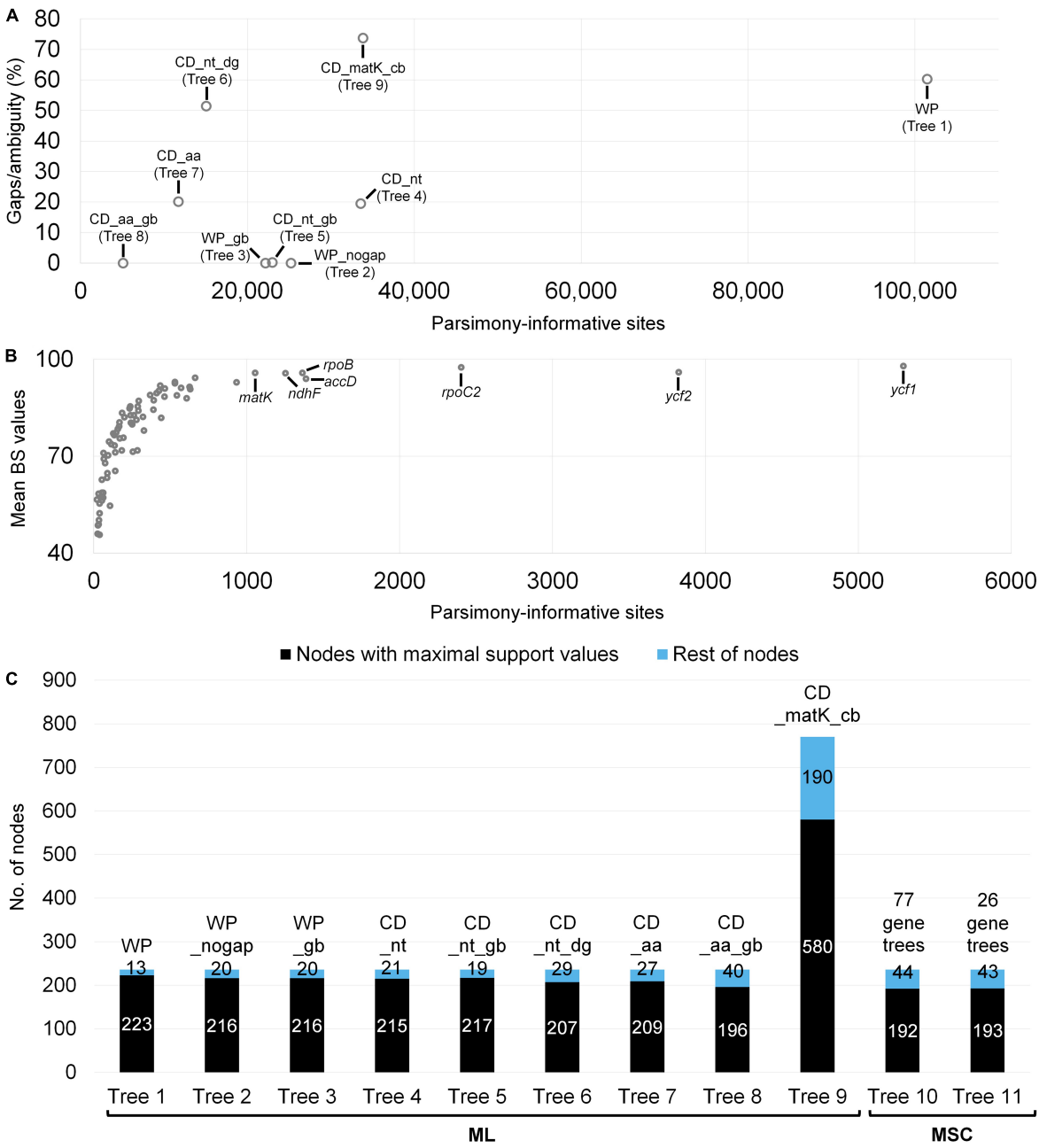

FIGURE 2 | Comparison of datasets used for phylogenetic analyses and their resolution in resulting trees. Descriptions for these datasets provided in **Table 1**. **(A)** A scatter plot showing differences in number of parsimony-informative sites (PIS) and gap/ambiguity among datasets for maximum likelihood (ML) analysis. **(B)** A scatter plot showing PIS and mean bootstrap values (BS) of 77 individual gene trees that were used for the multispecies coalescent (MSC) approach. The top seven protein-coding regions (including *matK*) with PIS (>1,000) are indicated. **(C)** A comparison of tree topology resolutions. Values within the histograms indicate the number of nodes. Maximal support values are BS = 100 or local posterior probability = 1.0.

to accommodate for poorly aligned sequences. While the lability of papilionoid plastome intergenic regions may provide a much greater number of PIS, their inclusion can introduce error in the alignment and infer spurious relationships.

In the remaining ML trees based on datasets without highly divergent intergenic regions, three controversial nodes connected by very short branches were revealed within a monophyletic group that included the Vataireoid, *Dermatophyllum*, Genistoid s.l., Dalbergioid s.l., and Baphieae + NPAAA clades (see **Figures 3–5**). The trees that were derived from analyses of translated AA sequences showed slightly better support. Even though MSC methods were suggested as an alternative way to explore plastome data (Gonçalves et al., 2019, 2020), this approach also left problematic nodes with low support values

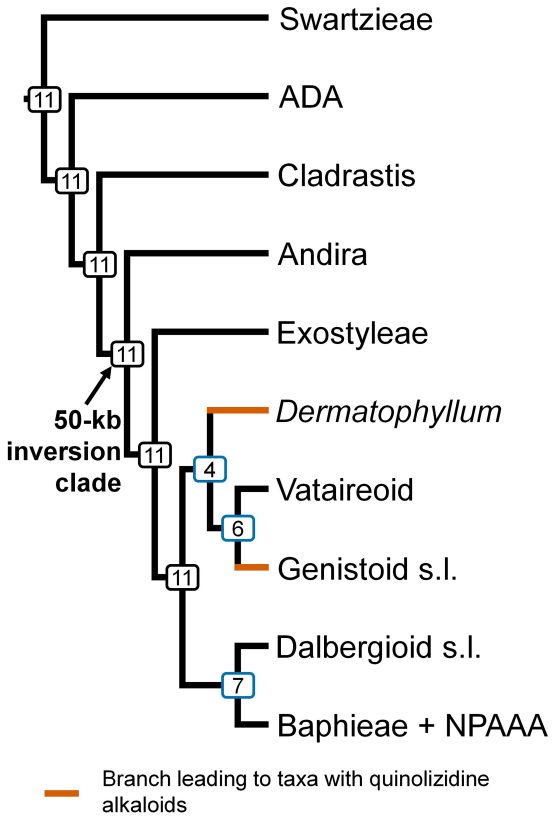

FIGURE 3 | The consensus of 11 phylogenetic trees of papilionoid legumes based on analyses of plastome-scale and *matK* sequence data. The number of topologies in which this clade was recovered in 11 trees is indicated at each node. Branches leading to taxa known to produce quinolizidine alkaloids are highlighted in brown. NPAAA: non-proteinogenic amino acid accumulating, ADA: Angylocalyceae, Dipterygeae, and Amburaneae, kb: kilobase, s.l.: sensu lato.

and showed a high frequency of alternative tree topologies. The source of conflicting signals among plastid loci can be diverse, including systematic and stochastic errors (Walker et al., 2019). The vast majority of plastid genes are short (see **Figure 2B** and **Supplementary Table 4**) and often yield poorly resolved individual gene trees as input for the MSC approach (Walker et al., 2019; Doyle, 2021). The species tree estimated from the trees based on the highly informative 26 CDS loci that exhibited strong phylogenetic signals also showed a similar level of alternative topologies (**Figure 4**). A likely explanation for this phylogenetic uncertainty is that the time interval between the divergences was too short to achieve resolution even with such phylogenetically informative loci. This makes it challenging to avoid stochastic errors in a zone of rapid divergence at deep nodes, such as the case within the 50-kb inversion clade of the papilionoid phylogeny. Thus, comparison with a phylogeny based on other genomic data (e.g., Zhao et al., 2021) would be desirable.

So far, several plastid phylogenomic analyses at several taxonomic levels of Fabaceae have been conducted (LPWG, 2017; Koenen et al., 2020a; Oyebanji et al., 2020; Zhang et al., 2020; Aecyo et al., 2021), but none of these has paid particular attention to increasing taxonomic diversity to fully resolve the deep nodes of Papilionoideae. Our taxon sampling included one to several representatives of most evolutionary key lineages except for a few small groups (e.g., Cardoso et al., 2013a; LPWG, 2017). In the following sections, phylogenetic relationships of papilionoid lineages will be discussed in a systematic context with other relevant molecular and non-molecular evidence. Overall, our analyses have provided well-resolved and strongly supported phylogenies, which are in agreement with topologies retrieved from previous plastid sequence-based phylogenies, and essentially validate the resolving power of *matK* sequences alone for reconstructing papilionoid phylogenies (e.g., Hu et al., 2000; Lavin et al., 2001; McMahon and Hufford, 2004; Cardoso et al., 2013a). Compared to a recent nuclear phylogenomic study (Zhao et al., 2021), there were several notable relationships with various nodal support values that were not retrieved from plastid data alone. Sexual recombination is restricted or absent in wild-type plastomes, suggesting possible involvement of reticulation/introgression events in the formation of some ancestral papilionoid lineages.

Swartzieae, ADA, and Cladrastis clades

The phylogenetic analyses in this study consistently resolved the early-diverging Swartzieae, ADA (Angylocalyceae, Dipterygeae, and Amburaneae), and Cladrastis clades, which collectively form a grade to the remaining Papilionoideae. Most genera within this grade were previously treated as members of tribes Sophoreae and Swartzieae by Polhill (1981b, 1994). In that sense, the grade comprising the Swartzieae, ADA, and Cladrastis clades, is reminiscent of Polhill's (1981b, **Figure 4**; 1994) traditional "supposed relationship of tribes," which recognized several groups of caesalpinioid-like genera among the largely tropical and subtropical Swartzieae and Sophoreae as transitional between subfamilies Caesalpinioideae and Papilionoideae. This group was considered to form a basal assemblage of the subfamily leading to the more derived papilionoid tribes. The taxonomic composition of this grade was partly recognized from early surveys for the absence/presence of a 50-kb inversion in legume plastomes (Doyle et al., 1996) and nodulation ability (Sprent, 2000).

Molecular phylogenetic analyses to date (Cardoso et al., 2012a, 2015; Ramos et al., 2016; LPWG, 2017; Zhang et al., 2020; Zhao et al., 2021) have reached a consensus that there are three main monophyletic groups that diverged early at the base of Papilionoideae. Early plastid sequence-based phylogenetic studies (Doyle et al., 1996, 1997; Pennington et al., 2001; Wojciechowski et al., 2004) suggested that groups of certain genera from the Swartzieae, Sophoreae, and Dipterygeae of Polhill's traditional classification (Polhill, 1981b, 1994) were outside of the 50-kb inversion clade in Papilionoideae. A Swartzieae-derived monophyletic group as the first diverging lineage sister to all Papilionoideae was initially recognized by Pennington et al. (2001), while Doyle et al. (1997) and Wojciechowski et al. (2004) showed groups of additional genera from these tribes among the early-diverging clades. A sister relationship between a clade comprising the genera *Cladrastis* Raf., *Pickeringia* Nutt. ex Torr. & A.Gray, and *Styphnolobium*

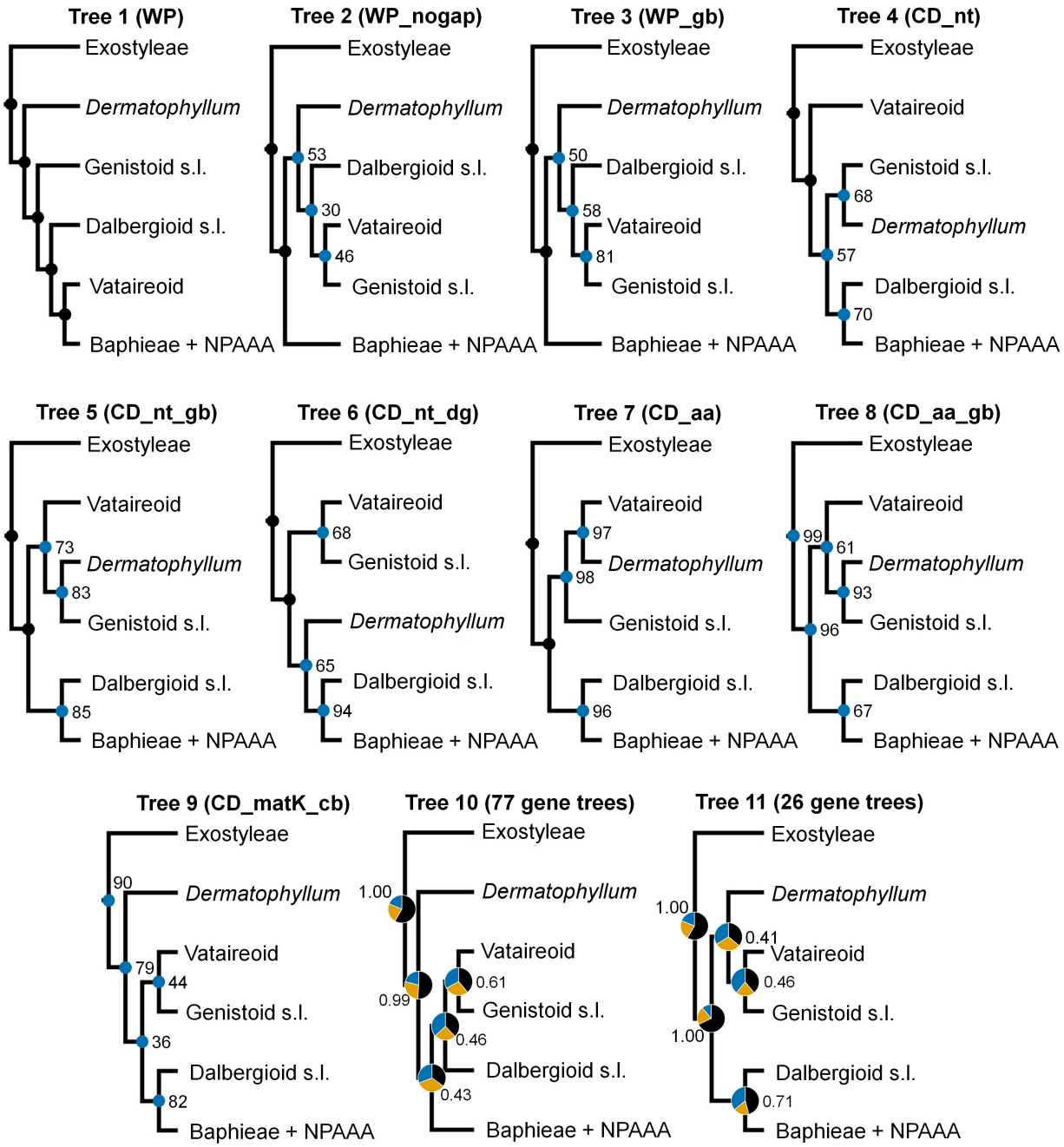

FIGURE 4 | Differences in topology and support values for the 11 trees with respect to resolving the phylogenetic relationships among the Vataireoid, *Dermatophyllum*, Genistoid s.l., Dalbergioid s.l., and Baphieae + NPAAA (non-proteinogenic amino acid accumulating) clades. Nodes with maximal bootstrap support (100) are denoted as black dots, while other nodes are marked with blue dots with values in nine trees (Trees 1–9). In Trees 10 and 11, based on the multispecies coalescence approach, pie charts representing quartet score values of topologies (black = main, orange = first alternative, blue = second alternative) and local posterior probability are presented at nodes.

Schott, and the more derived 50-kb inversion clade, was initially revealed by Wojciechowski et al. (2004). Subsequent studies by Cardoso et al. (2012a; 2013a; 2015), Wojciechowski (2013a), and Duan et al. (2019), with more comprehensive taxon sampling based upon analyses of both plastid *matK* and nuclear rDNA regions, further clarified the composition of the Swartzieae, ADA, and Cladrastis clades. Relationships among these three clades remain uncertain. Recent plastome-based phylogenetic analyses employing multi-locus (Ramos et al., 2016) and complete plastome (supported by all 11 trees, see **Figure 3**; Zhang et al., 2020) datasets retrieved the (Swartzieae,(ADA,(Cladrastis,50-kb inversion clade))) topology

while a recent nuclear phylogenomic study recovered the (ADA(Swartzieae,(Cladrastis,50-kb inversion clade))) topology (Zhao et al., 2021).

Given these results, the phylogenetic relationship of the Swartzieae and ADA clades remain equivocal. Both clades are highly diverse in floral structure (**Figure 1**) where almost all constituent genera display a particular floral architecture (e.g., Tucker, 1993, 2003b; Cardoso et al., 2013a; Leite et al., 2014, 2015; Prenner et al., 2015; Sinjushin, 2018). In addition to being distinguished by bilaterally or radially symmetrical apetalous flowers with a profusion of free stamens (e.g., *Cordyla*, *Swartzia* Schreb.; **Figure 1A**) or by single-petal flowers (e.g., *Amburana* Schwacke and Taub., *Swartzia*, *Trischidium* Tul.), representatives in these clades may have tiny radially symmetrical flowers measuring up to 3 mm long (*Myrocarpus* Allem., **Figure 1D**) to hardy, bat-pollinated flowers larger than 10 cm long (e.g., *Alexa* Moq.). Also, these clades include representatives with bilaterally symmetrical papilionate flowers that are unique in the Papilionoideae, either because of the wing-like, enlarged calyx lobes (e.g., *Dipteryx* Schreb., *Pterodon* Vogel; **Figure 1C**) or the fimbriate-glandular wing petals (e.g., *Petaladenium* Ducke). Such dramatic floral diversity in the early stages of the diversification history of the Papilionoideae indicates that resolving the relationships of the Swartzieae and ADA clades is fundamental, not just to better reconstruct the most likely ancestral form of the flower of the subfamily, but also to understand why their evolutionary history has been marked by profound deviations in flower morphology, from the typical papilionate architecture (as exhibited by those in **Figures 1F,I–K,M–O**).

Meso-Papilionoideae (50-kb Inversion Clade)

Large inversions in the plastome were first identified in legumes (Palmer and Thompson, 1982; Bruneau et al., 1990) and subsequently in other plant groups (e.g., Asteraceae, Jansen and Palmer, 1987; Poaceae, Doyle et al., 1992). Early evidence for a monophyletic group of papilionoids marked by this apparently unique, derived synapomorphy (Doyle et al., 1996) has been confirmed by subsequent molecular phylogenetic studies (e.g., Doyle et al., 1997; Pennington et al., 2001; Wojciechowski et al., 2004). The presence/absence of the 50-kb inversion in sampled taxa (Doyle et al., 1996) not only clarified the positions of several generic groups of both Sophoreae and Swartzieae at the base of papilionoids or among more derived groups (Polhill, 1981b), but also revealed the existence of a large monophyletic group of higher papilionoids defined by this molecular synapomorphy that includes the vast majority of the subfamily. Indeed, this large core papilionoid group comprises 98% of Papilionoideae based upon current species diversity estimates (LPWG, 2021). With this fact in mind, and consistent with the principles of phylogenetic nomenclature (Cantino and de Queiroz, 2006; Cantino et al., 2007), we suggest the adoption of "Meso-Papilionoideae" for the 50-kb inversion clade as defined previously by Wojciechowski (2013b). The 50-kb inversion serves as a synapomorphy for this clade, supported by nuclear (Zhao et al., 2021), mitochondrial (Choi et al., 2021), and plastid (supported by all 11 trees, see **Figure 3**; Zhang et al., 2020) phylogenomic studies.

The recent discovery of three species of *Sesbania*, nested within the Hologalegina clade (Robinioid + IRLC; **Figure 5**; Wojciechowski et al., 2000), with plastomes that appear to have completely reverted the 50-kb inversion (Lee et al., 2021) does not diminish the phylogenetic significance of this large inversion early in papilionoids. The phylogenetic distribution of this plastome structure within *Sesbania* or in its close relatives in the Robinioid clade remains elusive.

Meso-Papilionoideae includes the species-rich Genistoid s.l., Dalbergioid s.l., and Baphieae + NPAAA clades, the small Andira, Exostyleae and Vataireoid clades, as well as the phylogenetically unresolved genera *Dermatophyllum* and the African *Amphimas*. Relationships among and within these clades will be discussed below except for *Amphimas*, from which complete plastome data is not yet available, but which has been placed in the 50-kb inversion clade based on *matK* sequence data (**Figure 6**; Cardoso et al., 2015).

Andira and Exostyleae Clades

In our study, the Andira and Exostyleae clades were consistently (supported by all 11 trees, see **Figure 3**) resolved as successive sister groups to the monophyletic group comprising the remainder of the Meso-Papilionoideae (maximal support values from all three trees, see **Figure 5A**). The plastid phylogenomic study of Zhang et al. (2020), which included one taxon from each clade, also inferred the same topology. However, nuclear phylogenomic analysis (Zhao et al., 2021) placed *Andira inermis* (W. Wright) Kunth ex DC. sister to a clade of Baphieae + NPAAA, while Exostyleae was resolved as the first diverging lineage within Meso-Papilionoideae with maximum support values in all seven coalescent trees. The Andira clade (sensu Ramos et al., 2016), which includes three genera (*Aldina*, *Andira* Lam., and *Hymenolobium* Benth.), is expanded from Cardoso et al. (2012a, 2013a) to include *Aldina* based primarily on plastid data with some morphological similarities and shared ecological and distributional preference for Amazonian tropical rain forests. The Andira clade contains a mixture of taxa with radially symmetrical flowers with a profusion of exposed free stamens (*Aldina*) (see **Figure 1D**) and truly papilionate floral architecture (*Andira* and *Hymenolobium*). This was interpreted as an additional example of the common, interlaced phylogenetic distribution of the heterogeneous floral morphologies in early-diverging clades of Papilionoideae (Ramos et al., 2016). Whether the unexpected, more derived placement of *Andira* in Zhao et al. (2021) represents a signature of incongruence between plastid and nuclear genomes deserves further investigation.

Genistoid s.l. and *Dermatophyllum* Relationships, and Implications for the Evolutionary Distribution of Quinolizidine Alkaloids in Legumes

The sister relationship between the Genistoid s.l. and *Dermatophyllum* lineages was retrieved from three ML analyses,

where the highest support value (BS = 93) was found in Tree 8 (**Figure 4**). However, this relationship is equivocal because of conflicting alternative ML topologies based on analyses of the datasets derived from complete plastomes. In addition, the MSC tree estimations returned weakly supported relationships for these groups and inferred a high frequency of alternative topologies among individual CDS phylogenies.

Among legumes, QA production is restricted to most genistoid genera and to *Dermatophyllum* (Bisby et al., 1994; Kite and Pennington, 2003; Lee et al., 2013; Wink, 2013). This chemotaxonomic evidence has led to alternative hypotheses with respect to the controversial phylogenetic position of *Dermatophyllum*. One hypothesis suggests that the production of QAs is a derived characteristic defining a clade that includes the most recent common ancestor (MRCA) of the Genistoid s.l. and *Dermatophyllum* lineages, and thus forms strong evidence for a sister relationship of these two taxa (e.g., Cardoso et al., 2012a, 2013a; Kite et al., 2013; Lee et al., 2013). Alternatively, the genetic capacity for QA biosynthesis was established in the very early diversification of papilionoids but the genes remain silent (or were lost) in many descendant lineages (e.g., Wink et al., 2010). The Genistoid s.l. clade, as delimited in the phylogenetic analysis of *matK* sequences by Wojciechowski et al. (2004), included all known QA-accumulating genera except *Dermatophyllum* (as syn. *Calia* Terán and Berland.) and *Ormosia* Jacks. The Genistoid s.l. clade was subsequently expanded to accommodate the genus *Ormosia* and related genera of the Ormosieae, which has been consistently resolved as sister to the

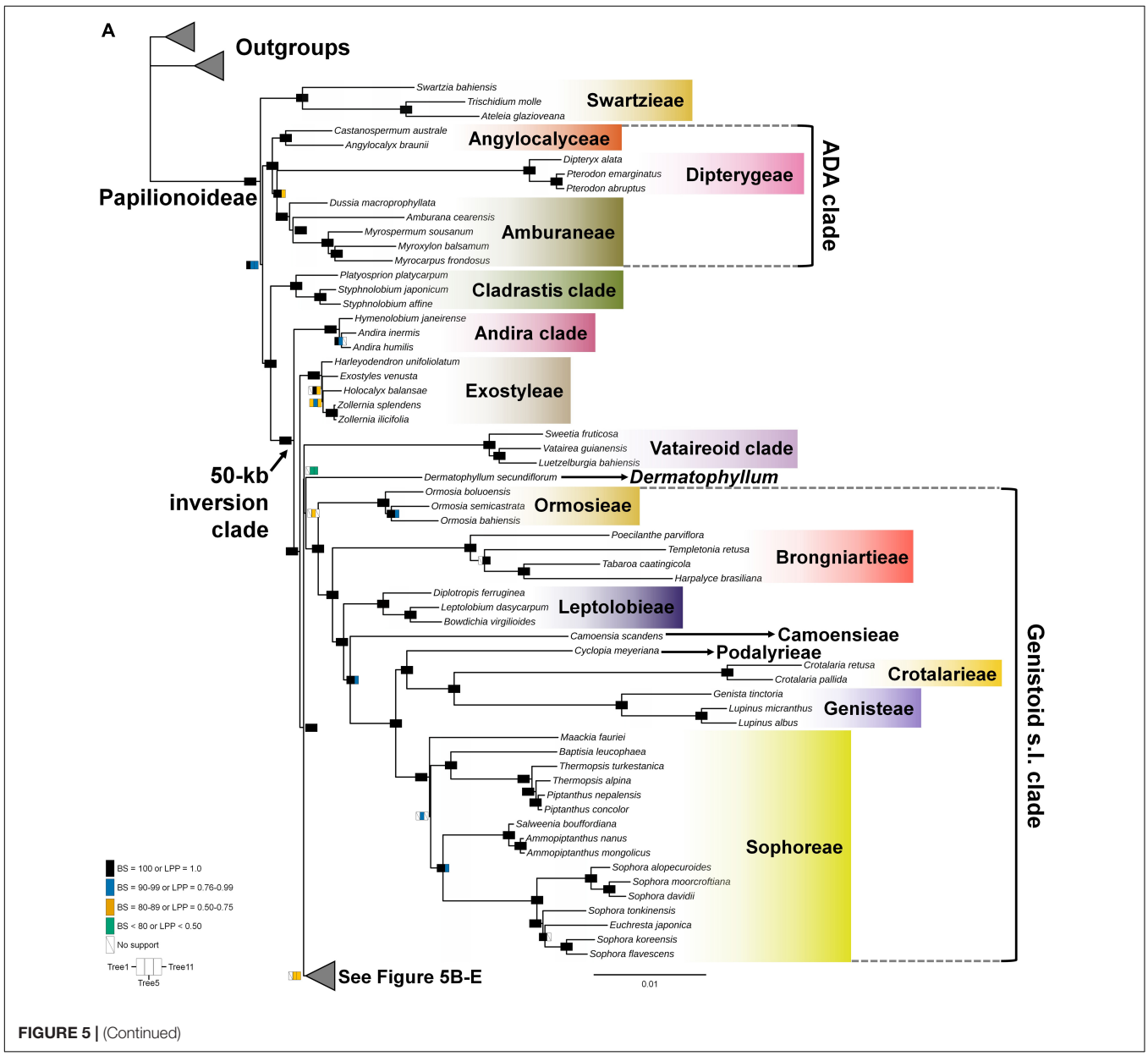

FIGURE 5 | (Continued)

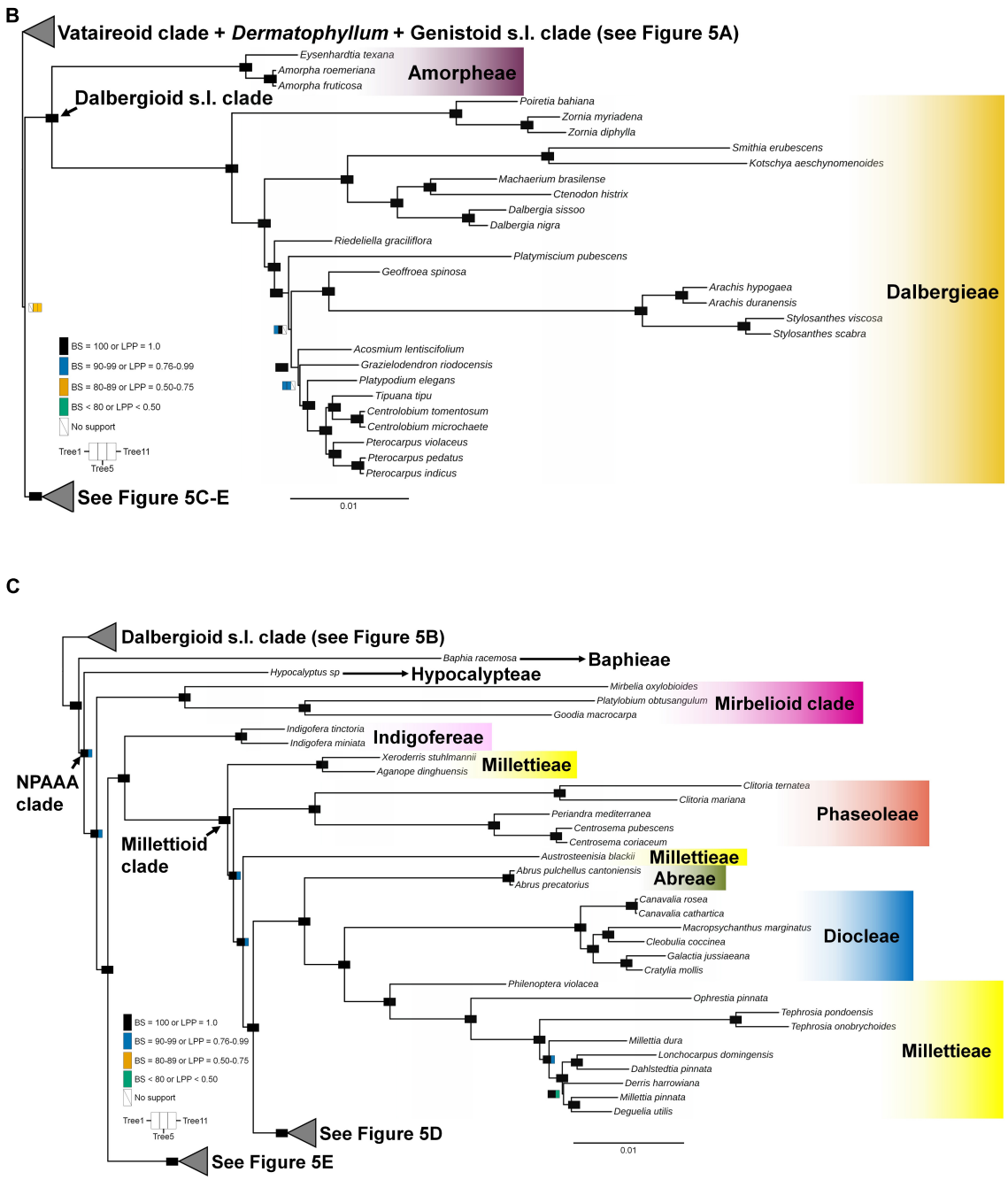

FIGURE 5 | (Continued)

rest of the clade (Cardoso et al., 2012a, 2013a), suggesting that the only remaining QA-producing genus *Dermatophyllum* was likely sister to the Genistoid s.l. clade. The scenario of Wink et al. (2010) was based on an *rbcL* gene phylogeny that showed an early divergence of *Dermatophyllum* within a monophyletic group containing *Myroxylon* L. (Amburaneae) that does not produce QAs and is not a member of Meso-Papilionoideae. However, our study (**Figures 3–5**) and others (Zhang et al., 2020) with broad taxon sampling of papilionoid legumes resolved *Dermatophyllum* as a more derived lineage in the Meso-Papilionoideae. Wink et al. (2010) suggested that QA production was one of many herbivore defense strategies rendering it dispensable, as exemplified by loss or extreme reduction of QA production in some species within genistoid genera such as *Crotalaria* L., *Lotononis* (DC.) Eckl. and Zeyh., *Ulex* L., *Calicotome* Link, and *Spartocytisus* Webb and Berthel.

In light of our results, modified ancient gain and loss scenarios regarding QA production can be postulated according to the

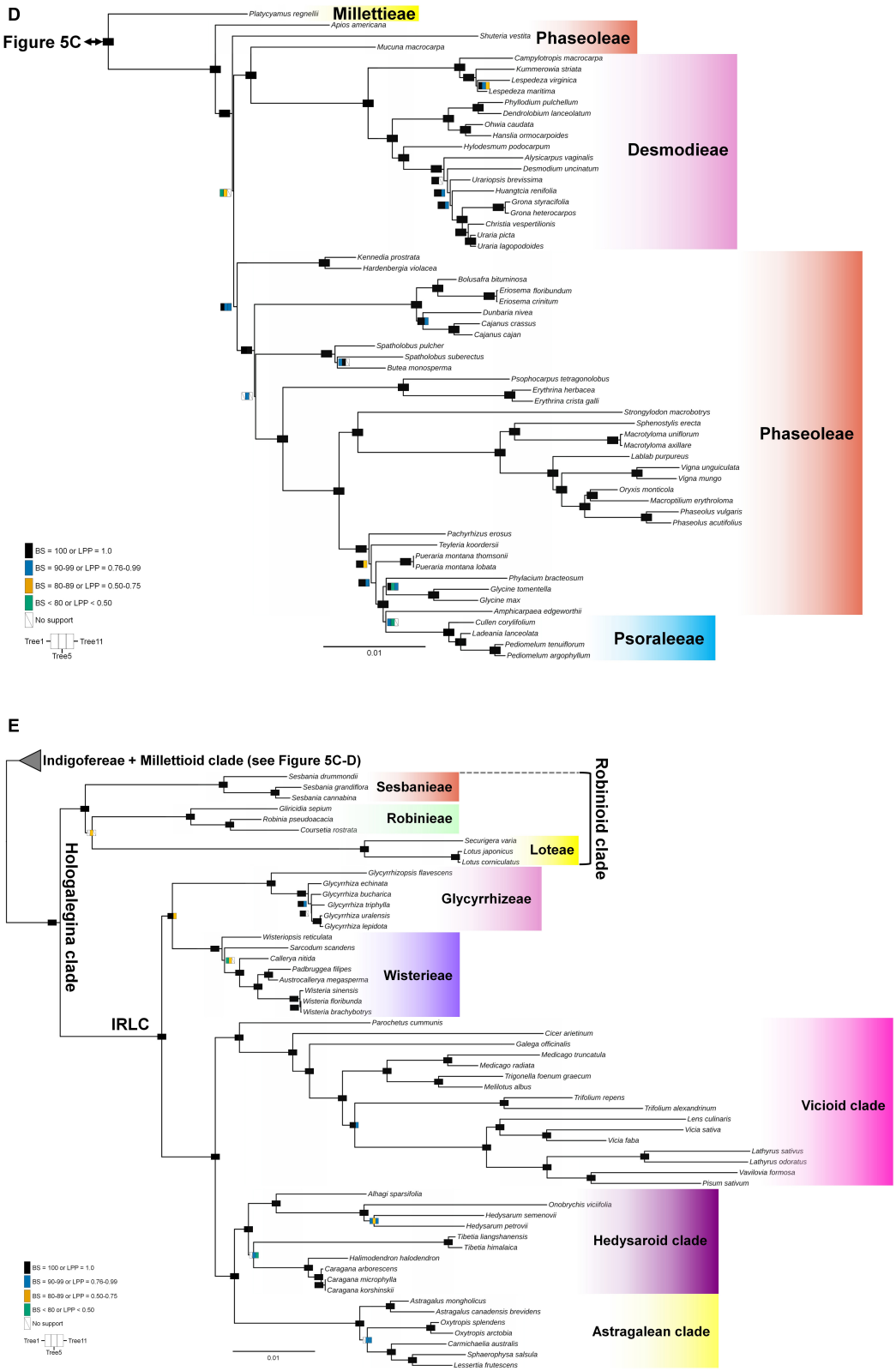

FIGURE 5 | A maximum-likelihood tree based on a concatenated dataset of 77 plastome coding regions with Gblocks trimming (See **Table 1**). The figure is separated into five panels with a focus on the major clades. **(A)** Swartzieae, ADA (Angylocalyceae, Dipterygeae, and Amburaneae), Cladrastis, Andira, Exostyleae,

(Continued)

FIGURE 5 | Vataireoid, and Genistoid s.l. clades; **(B)** Dalbergioid s.l. clade; **(C)** Baphieae, Hypocalypteae, Mirbelioid, Indigofereae, Millettioid (part 1) clades; **(D)** Millettioid (part 2) clade and **(E)** Robinioid clade and inverted repeat-lacking clade (IRLC). Supporting values of three trees are visualized at nodes as rectangles. The rectangles are subdivided into three columns that correspond to Tree 1 (left, WP), Tree 5 (middle, CD_nt_gb), and Tree 11 (right, 26 gene trees). Colors in each column represent support values. Black [bootstrap support value (BS) = 100 or local posterior probability (LPP) = 1], Blue [BS (90–99) or LPP (0.76–0.99)], Orange [BS (80–89) or LPP (0.50–0.75)], Bluish green (BS < 80 or LPP 0.50). Scale indicates number of nucleotide substitutions per site. The colors for clades are not related to their support values. We generally followed the color scheme for early-branching papilionoid clades of Figures 2–5 in Cardoso et al. (2013a).

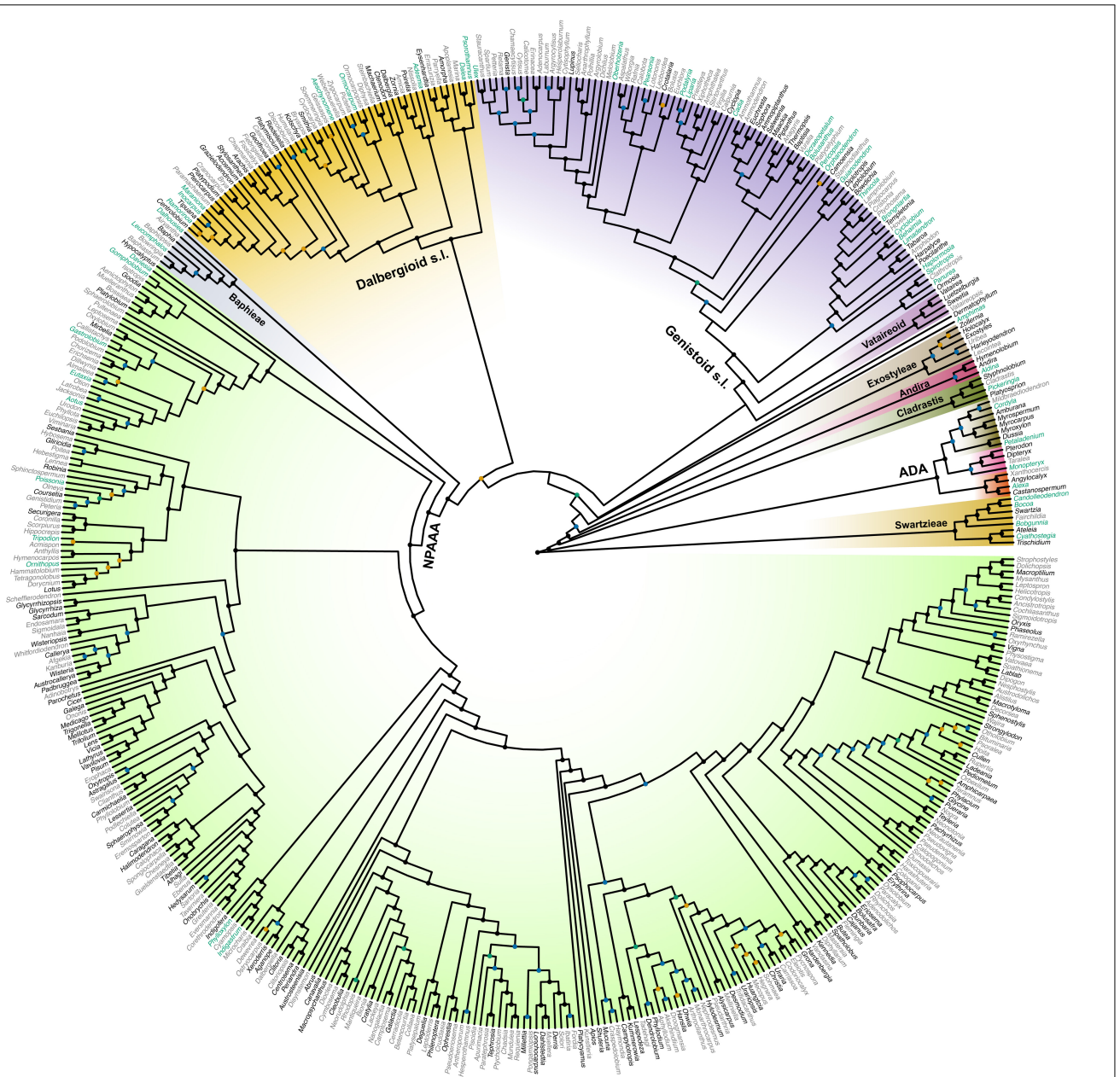

FIGURE 6 | A genus-level phylogeny of Papilionoideae as produced by the maximum likelihood analysis of the dataset 9 (CD_matK_cb), which combined genera with fully sequenced plastomes and the expanded sampling of genera with *matK* sequence data only. Black nodes are supported by bootstrap values (BS) = 100 and blue nodes by BS = 90–99, while orange nodes are supported by BS = 80–89 and cyan nodes by BS = 70–79. Genera highlighted in black have already been fully sequenced for plastomes. The 50 genera highlighted in cyan are either still phylogenetically enigmatic, are morphological key genera, or represent monospecific lineages, all of which should be the focus of future plastome research. The colors for clades are not related to their support values. We followed the color scheme for papilionoid clades of Figure 1 in Cardoso et al. (2013a). NPAAA: non-proteinogenic amino acid accumulating, ADA: Angylocalyceae, Dipterygeae, and Amburaneae.

alternative phylogenetic positions of *Dermatophyllum* (supported by BS value > 90) relative to the Genistoid s.l., Vataireoid, Dalbergioid s.l., and Baphieae + NPAAA clades. One scenario posits that the QA biosynthesis pathway was present at least since the MRCA of all five lineages but was lost after the successive divergence of the *Dermatophyllum* and the Genistoid s.l. clades (**Figure 4**, Tree 1). Alternatively, QA production is ancestral in only three lineages (*Dermatophyllum*, Genistoid s.l. and Vataireoid clades) but was lost early in the diversification of the Vataireoid clade. Our study does not confidently identify the position of *Dermatophyllum*, but it reduces the number of possible solutions by resolving other early-diverging relationships within Meso-Papilionoideae. Nevertheless, the *Dermatophyllum*-Genistoid s.l. sister hypothesis remains the most parsimonious (single-step, excluding recent loss events in the genistoids) scenario. In order to shed light on the evolutionary pathway(s) leading to the biosynthesis of QAs, resolving the relationship of *Dermatophyllum* with respect to the remaining genistoids is a high priority in future nuclear-based phylogenomics of the early-diverging Papilionoideae.

Vataireoid Clade

The exclusively neotropical Vataireoid clade includes just 28 species, and is another example of an early-diverging papilionoid lineage outside the large NPAAA clade with heterogeneous floral morphology (Cardoso et al., 2013b), similar to the Andira clade (sensu Ramos et al., 2016). Monophyly of this morphologically heterogeneous group was highly supported in phylogenetic analyses based on single plastid genes or a few combined nuclear and plastid genes (Cardoso et al., 2012a, 2013a,b) to genome-scale data (**Figure 5A**; Zhao et al., 2021). There was no strongly supported sister relationship hypothesis for this clade except Zhao et al. (2021). The close affinity of the vataireoids to Dalbergieae based on a single-seeded samaroid fruit morphology was suggested by Lima (1980; 1982; 1990), but molecular systematic studies have not supported this relationship. The phylogenetic position of the clade is inconclusive based on our analyses, but a nuclear phylogenomic study (Zhao et al., 2021) resolved the clade as sister to a monophyletic group (including Dalbergioid s.l., Genistoid s.l., Andira, and Baphieae + NPAAA clades) that includes all known taxa with the ability to nodulate within Meso-Papilionoideae (Sprent et al., 2017; Ardley and Sprent, 2021).

Dalbergioid s.l. Clade

The Dalbergioid s.l. clade includes monophyletic groups of the pantropical Dalbergieae and the predominantly North American temperate Amorpheae (Wojciechowski et al., 2004). In our study, this clade most frequently grouped with the NPAAA + Baphieae clade, albeit with alternative positions and low support values (**Figures 3**, **4**). A sister group relationship of the Andira clade with the Dalbergioid s.l. clade was once weakly supported (Wojciechowski et al., 2004), but further support for that relationship is lacking. A nuclear phylogenomic study (Zhao et al., 2021) grouped the Dalbergioid s.l. clade with the Genistoid s.l. clade but with weak support (highest BS value from seven coalescent trees was 77). As such, the sister relationship of the dalbergioids within Meso-Papilionoideae is still unclear.

While both nuclear phylogenomic (Zhao et al., 2021) and our plastid phylogenomic (**Figures 3–5**) analyses have failed to clarify the relationships of the Dalbergioid s.l. clade, they concur with previous comprehensively sampled *matK* phylogenies in strongly supporting the monophyly and interrelationships of three main subclades within the dalbergioids: the *Adesmia*, *Dalbergia*, and *Pterocarpus* clades (Lavin et al., 2001; Wojciechowski et al., 2004; Cardoso et al., 2012a, 2013a). Additionally, by resolving the radially symmetrical flowered genera *Riedeliella* Harms and *Acosmium* Schott in isolated positions within Dalbergieae, our analyses based on complete plastomes demonstrated yet again how the independent evolution of non-papilionate floral architecture has been so recurrent among the early-branching papilionoids (Lavin et al., 2001; Cardoso et al., 2012a,b).

Baphieae and Non-protein Amino Acid Accumulating Clades

A single origin of non-protein amino acid biosynthesis in papilionoids was hypothesized by Bell (1981) because canavanine, a close analog of arginine, was almost mutually exclusive of alkaloid accumulation, and restricted to 16 closely related papilionoid tribes. A monophyletic group containing all known taxa producing non-protein amino acids, the NPAAA clade (Cardoso et al., 2012b), is supported by all recent phylogenetic analyses (LPWG, 2017; Koenen et al., 2020b, 2021; Zhang et al., 2020; Choi et al., 2021; Zhao et al., 2021). This clade includes several of the most species-rich (ca. 11,000 spp.) and rapidly evolving legume lineages (Cardoso et al., 2013a; Koenen et al., 2013), as well as the largest flowering plant genus *Astragalus* L. (ca. 3000 spp.) and the most agriculturally important culinary pulses such as common beans, peas, lentils, and soybeans (Gepts et al., 2005). The clade also stands out among the main Papilionoideae lineages with respect to the almost universal evolutionary canalization of the specialized, strongly bilaterally symmetrical papilionate floral architecture. Thus far, there are no examples of reversion to radial floral symmetry or profusion of free stamens as found in earlier branching clades (Lewis et al., 2005; Cardoso et al., 2013a).

The NPAAA clade is sister to the small (c. 60 spp.), predominantly African Baphieae clade, which contains genera with both radially and bilaterally symmetrical flowers. Evolutionary transitions between polysymmetry and monosymmetry in floral architecture are common in angiosperms (Endress, 1996, 1999), such that the emergence of the core eudicots, a clade with more than 200,000 species, coincides with the fixation of pentamerous flowers, whorl organization, and a perianth often differentiated into sepals and petals (Specht and Bartlett, 2009). Such changes in floral architecture may have led to the recurrent evolution of bilateral symmetry (zygomorphy) from polysymmetric-flowered ancestors across angiosperms. This shift coincides with the co-diversification of megadiverse families and specialized pollinating insects (Cardinal and Danforth, 2013). These factors may have contributed to increased speciation rates

(Sargent, 2004; Davis et al., 2014) and monosymmetry as a key innovation during the radiation of angiosperms (Sanderson and Donoghue, 1994; Bond and Opell, 1998). Likewise, the evolutionary maintenance of the papilionate flower may have sparked the explosive diversification of the large NPAAA clade, which includes almost 70% of both specific and generic diversity in Papilionoideae (Cardoso et al., 2012a, 2013b).

Here, two large clades within the NPAAA clade have been confirmed in agreement with all previous molecular phylogenetic studies (**Figures 5C–E**; Wojciechowski et al., 2004; Cardoso et al., 2013a; Queiroz et al., 2015; LPWG, 2017; Zhao et al., 2021): the Indigofereae + Millettioid and Hologalegina (Robinioid + IRLC) clades. Most of the relationships among the main lineages within the Millettioid clade concur with previous studies (**Figures 5C,D**; Queiroz et al., 2015; Zhao et al., 2021), however, a discordance was observed in the sister relationship of *Mucuna* Adans. (Phaseoleae) and Desmodieae. *Mucuna* was supported as the sister to Desmodieae in plastid phylogenies (Doyle et al., 2000; Kajita et al., 2001; Stefanović et al., 2009; Jin et al., 2019). This sister relationship is also marked by the shared loss of the plastid *rpl2* intron, which appears to have been lost only a few times in legumes (Doyle et al., 1995; Bailey et al., 1997; Lai et al., 1997; Jin et al., 2019). While our analyses of complete plastome data showed *Mucuna* as sister (maximal support values from all three trees, see **Figure 5D**), or part of a sister clade (including *Craspedolobium* Harms and *Haymondia* A.N.Egan & B.Pan) to Desmodieae (**Figure 6**), Zhao et al.'s (2021) nuclear phylogenomic analyses resolved *Mucuna* as sister to *Cochlianthus* Benth., combining a Mucuna-Cochlianthus clade as sister to *Apios* Fabr., whereas *Craspedolobium* and *Haymondia* were successive sister groups of Desmodieae. Lackey (1981) considered the genera *Apios*, *Cochlianthus*, and *Mucuna* to be members of an artificial amalgamation that defined the Phaseoleae subtribe Erythrininae, but considered *Apios* and *Cochlianthus* to be a natural grouping or even congeneric, making the sister relationship between *Cochlianthus* and *Mucuna* largely unexpected. However, the key feature of *Haymondia* (i.e., an explosive flower tripping mechanism involving the upward movement of the reproductive column that remains fully reflexed from the wing and keel petals, and touching the standard petal; Egan and Pan, 2015) can also be found in some genera of Phaseoleae (*Apios*, *Cochlianthus*, and *Mucuna*) and Desmodieae (except the Lespedeza group) within the millettioids (Schrire et al., 2009). Due to its phylogenetically scattered distribution that includes Indigofereae and *Medicago* L. within the NPAAA clade, as well as in the genistoids *Genista* L., *Harpalyce* Moc. and Sessé ex DC., *Spartium* L., and *Ulex* L., and the dalbergioid *Brya* P. Browne (Arroyo, 1981; Lavin et al., 2001; Schrire et al., 2009), convergent evolution of this pollination-related morphological feature cannot be ruled out. The presence of a cryptic, shared morphological feature between two distantly related clades, including *Apios* and Desmodieae and the highly supported, yet conflicting positions of *Mucuna* in nuclear and plastid phylogenies supports putative ancestral hybridization in these groups (Egan et al., 2016).

The Hologalegina clade was first designated by Wojciechowski et al. (2000) to comprise the traditionally recognized "temperate herbaceous tribes" of Polhill (1981b, 1994). This large monophyletic group includes two main, well supported subclades; Robinoid (Sesbanieae, Loteae, and Robinieae) and the IRLC. Even though some relationships within the Robinioid clade and IRLC need further clarification, the monophyly of each group has been consistently supported in plastid-based phylogenies (maximal support values from all three trees, see **Figure 5E**; Wojciechowski et al., 2004; Lavin et al., 2005). The nuclear phylogenomics study of Zhao et al. (2021) did not support the monophyly of the Robinioids and resolved Sesbanieae + Loteae as sister to the IRLC, albeit with low support values (4 out of 7 trees > BS 70). Indeed, Loteae was once regarded as closer to members of the IRLC than to Robinieae, based on similarities in vegetative morphology, growth habit, and distribution (Polhill, 1981b, 1994). Similarly, the early nuclear rDNA-ITS-based phylogeny of Hu et al. (2000) and recent mitogenome-based phylogeny of Choi et al. (2021), with limited taxon sampling (without Sesbanieae), resolves *Lotus* L. as sister to the IRLC.

Within the Vicioid clade (IRLC), five genera of Trifolieae (*Medicago*, *Melilotus* (L.) Mill., *Ononis* L., *Trifolium* L., *Trigonella* L., excluding *Parochetus* Buch.-Ham. ex D.Don), have been resolved as paraphyletic, placing *Trifolium* as sister to the monophyletic Fabeae in plastid phylogenies (Wojciechowski et al., 2000, 2004; Steele and Wojciechowski, 2003; Schaefer et al., 2012). Our study also showed the sister relationship of *Trifolium* to Fabeae (maximal support values from two of three trees, see **Figure 5E**). However, the nuclear phylogenomic study resolved *Trifolium* as sister to a clade composed of *Medicago*, *Melilotus*, and *Trigonella* with maximal support values from all seven coalescent trees (Zhao et al., 2021). Similarly, but with very limited sampling, a mitochondrial phylogenomic study (Choi et al., 2021) also supported a monophyletic group based on *Medicago* and *Trifolium* as sister to *Vicia* (Fabeae). This grouping of four Trifolieae genera (*Medicago*, *Melilotus*, *Trifolium*, and *Trigonella*) is also marked by the loss of mitochondrial *rps1* due to its functional, intracellular gene transfer to the nuclear genome (Hazle and Bonen, 2007; Choi et al., 2020b). In the case of Trifolieae, in which the nuclear and mitochondrial data produce topologies that agree closely with a classification based on gross morphology while the plastid data does not, a plastid capture scenario is worth considering.

Phylogenies based on plastid and mitochondrial genomes can produce different topologies for a lineage because of biparental organelle inheritance together with cytonuclear incompatibility, as exemplified in the IRLC genus *Pisum* L. (Bogdanova et al., 2021). Many IRLC taxa share the potential for biparental plastid inheritance (Corriveau and Coleman, 1988; Zhang et al., 2003). That a limited number of taxa have been tested for potential paternal transmission of organelle genomes and that the mitogenome-based phylogeny showed an alternative topology with regard to *Lotus* and *Trifolium* (Choi et al., 2021) warrants further investigation on the mode of organelle inheritance in Hologalegina. Within the IRLC, plastid capture scenarios have been suggested for various lineages based on conflicting results between nuclear rDNA and plastid data-based phylogenies (e.g., Ellison et al., 2006; Duan et al., 2021).

Discordant taxon sampling across the three genomic datasets and the topological conflicts, which may have originated from the complex evolutionary behavior of repetitive nuclear rDNA (Álvarez and Wendel, 2003), hinder the detailed examination of those scenarios.

Future Research Directions in Plastid Phylogenomics of the Papilionoid Legumes

Our study provides well resolved and strongly supported plastid phylogenies for the papilionoid legumes. Nevertheless, there remain nodes lacking phylogenetic resolution. The inclusion of non-CDS regions of plastomes could add more phylogenetic signal but at the potential expense of introducing a great amount of homoplasy, which could either decrease phylogenetic support or introduce bias toward unreliable relationships. There are further limitations on the use of plastome data despite that they have served well as a fundamental source of phylogenetic information. This information has helped us to address complex evolutionary issues in the papilionoid legumes, such as the origin of their extremely diverse floral architectures and secondary metabolites. The ancient, rapid diversification and reticulation/introgression detected in plastome-based Papilionoideae phylogenies warrant the incorporation of nuclear and mitochondrial data from concordant, broad taxon sampling. Perception and reconciling of conflicting phylogenetic signals within and between the three genomes are in their primary stage, and more complex evolutionary patterns may be revealed from the total genomic loci across all levels of subdivision in the Papilionoideae. In Hologalegina in particular, where the transition of maternal to biparental plastid inheritance as well as reticulation/introgression(s) likely occurred, more conflicts are expected.

Closing the sampling gaps within and between *matK*-only and complete plastome datasets could shed light on future sampling directions in Papilionoideae phylogenomics. Since Cardoso et al. (2013a) estimated the number of early-branching papilionoid genera that were not sampled for *matK* sequence data (52 out of the 196), there has been great progress in filling the sampling gap by virtue of individual locus or whole plastome sequences (e.g., Cardoso et al., 2015, 2017; Swanepoel et al., 2015; LPWG, 2017; Queiroz et al., 2017; Zhang et al., 2020). Our *matK*-only + plastome combined data set (**Supplementary Table 5**) includes 39 genera that were previously unsampled by Cardoso et al. (2013a), however, 12 genera (one is synonymized, see **Supplementary Table 6**) still remain to be fully sequenced. Moreover, many genera are still not well resolved in phylogenies based on *matK*-only data, such as the African tree genus *Amphimas* (Cardoso et al., 2012a, 2013a, 2015). Further plastome research should focus on examining the positions of phylogenetically unresolved genera, as well as those that are morphological key groups or represent monospecific lineages (**Figure 6**). Next generation sequencing applied to museomics (Bakker et al., 2016; Johnson et al., 2019) has proven to be a feasible approach to collecting genomic data from taxa for which fresh tissues are unavailable.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: GenBank with accessions MZ725323, OL672849-OL672886. All sequence alignments and trees that were generated from this study are submitted to Dryad (https://doi.org/10.5061/dryad.sf7m0cg7m).

AUTHOR CONTRIBUTIONS

MW, TR, RJ, DC, and I-SC: conception and experimental design. MW, RJ, and TR: acquisition of funds. DC, RJ, TR, I-SC, HL, LQ, and MW: field collections. I-SC, CL, and TR: nucleic acid isolation. CL and I-SC: plastome assembly and annotation. I-SC, DC, and MW: DNA alignments and phylogenetic analyses and data interpretation. I-SC and DC: production of figures and tables. I-SC, DC, MW, TR, and RJ: writing and revision of the manuscript. All authors read and commented on the manuscript.

FUNDING

This work was supported by grants from the National Science Foundation (DEB-1853010 and DEB-1853024) to MW, RJ, and TR, the Texas Ecological Laboratory (EcoLab) to RJ, TR, and I-SC, the Sidney F. and Doris Blake Professorship in Systematic Botany to RJ, the CNPq (Research Productivity Fellowship no. 308244/2018-4; Universal no. 422325/2018-0), FAPESB (Universal no. APP0037/2016), and UFBA PROQUAD program to DC.

ACKNOWLEDGMENTS

We thank the TEX-LL, HUEFS, and RB herbaria for voucher deposition, the Desert Legume Program at the University of Arizona for seeds, and Alessandra Schnadelbach, Hedina Basile, Henrique Batalha Filho, and Paula Ristow for their lab support at UFBA. We also thank George Yatskievych (TEX/LL) for arranging a formal Material Transfer Agreement (Decree number 8772) under the SisGen Cadastro RDC6BE9, which facilitated research activities between our institutions. Finally, we wish to thank the two reviewers and the editor for their many helpful comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.823190/full#supplementary-material

REFERENCES

Aecyo, P., Marques, A., Huettel, B., Silva, A., Esposito, T., Ribeiro, E., et al. (2021). Plastome evolution in the Caesalpinia group (Leguminosae) and its application in phylogenomics and populations genetics. *Planta* 254:27. doi: 10.1007/s00425-021-03655-8

Álvarez, I., and Wendel, J. F. (2003). Ribosomal ITS sequences and plant phylogenetic inference. *Mol. Phylogenet. Evol.* 29, 417–434. doi: 10.1016/S1055-7903(03)00208-2

Antonelli, A., Clarkson, J. J., Kainulainen, K., Maurin, O., Brewer, G. E., Davis, A. P., et al. (2021). Settling a family feud: a high-level phylogenomic framework for the Gentianales based on 353 nuclear genes and partial plastomes. *Am. J. Bot.* 108, 1143–1165. doi: 10.1002/ajb2.1697

Ardley, J., and Sprent, J. (2021). Evolution and biogeography of actinorhizal plants and legumes: a comparison. *J. Ecol.* 109, 1098–1121. doi: 10.1111/1365-2745.13600

Arroyo, M. T. K. (1981). "Breeding systems and pollination biology in Leguminosae," in *Advances in Legume Systematics, Part2*, eds R. M. Polhill and P. H. Raven (Richmond, UK: Royal Botanic Gardens, Kew), 723–770.

Bailey, C. D., Doyle, J. J., Kajita, T., Nemoto, T., and Ohashi, H. (1997). The chloroplast *rpl2* intron and ORF184 as phylogenetic markers in the legume tribe Desmodieae. *Syst. Bot.* 22, 133–138. doi: 10.2307/2419681

Bakker, F. T., Lei, D., Yu, J., Mohammadin, S., Wei, Z., Van de Kerke, S., et al. (2016). Herbarium genomics: plastome sequence assembly from a range of herbarium specimens using an iterative organelle genome assembly pipeline. *Biol. J. Linn. Soc.* 117, 33–43. doi: 10.1111/bij.12642

Bell, E. A. (1981). "Non-protein amino acids in the Leguminosae," in *Advances in Legume Systematics*, *Part 2*, eds R. M. Polhill and P. H. Raven (Richmond, UK: Royal Botanic Gardens, Kew), 489–499.

Bisby, F. A., Buckingham, J., and Harborne, J. B. (eds). (1994). *Phytochemical Dictionary of the Leguminosae*. London: Chapman and Hall.

Bogdanova, V. S., Shatskaya, N. V., Mglinets, A. V., Kosterin, O. E., and Vasiliev, G. V. (2021). Discordant evolution of organellar genomes in peas (*Pisum* L.). *Mol. Phylogenet. Evol.* 160:107136. doi: 10.1016/j.ympev.2021.107136

Bond, J. E., and Opell, B. D. (1998). Testing adaptive radiation and key innovation hypotheses in spiders. *Evolution* 52, 403–414. doi: 10.1111/j.1558-5646.1998.tb01641.x

Bruneau, A., Doyle, J. J., and Palmer, J. D. (1990). A chloroplast DNA inversion as a subtribal character in the Phaseoleae (Leguminosae). *Syst. Bot.* 15, 378–386. doi: 10.2307/2419351

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

Cai, Z., Guisinger, M., Kim, H.-G., Ruck, E., Blazier, J. C., McMurtry, V., et al. (2008). Extensive reorganization of the plastid genome of *Trifolium subterraneum* (Fabaceae) is associated with numerous repeated sequences and novel DNA insertions. *J. Mol. Evol.* 67, 696–704. doi: 10.1007/s00239-008-9180-7

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

Cantino, P. D., and de Queiroz, K. (2006). *PhyloCode: a Phylogenetic Code of Biological Nomenclature. Version 3a*. Available online at: http://www.phylocode.org (accessed September 15, 2021).

Cantino, P. D., Doyle, J. A., Graham, S. W., Judd, W. S., Olmstead, R. G., Soltis, D. E., et al. (2007). Towards a phylogenetic nomenclature of Tracheophyta. *Taxon* 56, E1–E44. doi: 10.1002/tax.563001

Cardinal, S., and Danforth, B. N. (2013). Bees diversified in the age of eudicots. *Proc. Royal Soc. B* 280:20122686. doi: 10.1098/rspb.2012.2686

Cardoso, D., Harris, D. J., Wieringa, J. J., São-Mateus, W. M. B., Batalha-Filho, H., Torke, B. M., et al. (2017). A molecular-dated phylogeny and biogeography of the monotypic legume genus *Haplormosia*, a missing African branch of the otherwise American-Australian Brongniartieae clade. *Mol. Phylogenet. Evol.* 107, 431–442. doi: 10.1016/j.ympev.2016.12.012

Cardoso, D., Lima, H. C., Rodrigues, R. S., Queiroz, L. P., Pennington, R. T., and Lavin, M. (2012a). The realignment of *Acosmium* sensu stricto with the Dalbergioid clade (Leguminosae: papilionoideae) reveals a proneness for independent evolution of radial floral symmetry among early-branching papilionoid legumes. *Taxon* 61, 1057–1073. doi: 10.1002/tax.615011

Cardoso, D., Queiroz, L. P., Pennington, R. T., Lima, H. C., Fonty, E., Wojciechowski, M. F., et al. (2012b). 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

Cardoso, D., Pennington, R. T., Queiroz, L. P., Boatwright, J. S., Van Wyk, B.-E., Wojciechowski, M. F., et al. (2013a). Reconstructing the deep-branching relationships of the papilionoid legumes. *S. Afr. J. Bot.* 89, 58–75. doi: 10.1016/j.sajb.2013.05.001

Cardoso, D., Queiroz, L. P., Lima, H. C., Suganuma, E., Van den Berg, C., and Lavin, M. (2013b). A molecular phylogeny of the vataireoid legumes underscores floral evolvability that is general to many early-branching papilionoid lineages. *Am. J. Bot.* 100, 403–421. doi: 10.3732/ajb.1200276

Cardoso, D., São-Mateus, W. M., Cruz, D. T., Zartman, C. E., Komura, D. L., Kite, G., et al. (2015). Filling in the gaps of the papilionoid legume phylogeny: the enigmatic Amazonian genus *Petaladenium* is a new branch of the early-diverging Amburaneae clade. *Mol. Phylogenet. Evol.* 84, 112–124. doi: 10.1016/j.ympev.2014.12.015

Castellanos, C., Steeves, R., Lewis, G. P., and Bruneau, A. (2017). A settled sub-family for the orphan tree: the phylogenetic position of the endemic Colombian genus *Orphanodendron* in the Leguminosae. *Brittonia* 69, 62–70. doi: 10.1007/s12228-016-9451-3

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

Charboneau, J. L. M., Cronn, R. C., Liston, A., Wojciechowski, M. F., and Sanderson, M. J. (2021). Plastome structural evolution and homoplastic inversions in Neo-Astragalus (Fabaceae). *Genome. Biol. Evol.* 13:evab215. doi: 10.1093/gbe/evab215

Choi, I.-S., and Choi, B.-H. (2017). The distinct plastid genome structure of *Maackia fauriei* (Fabaceae: papilionoideae) and its systematic implications for genistoids and tribe Sophoreae. *PLoS One* 12:e0173766. doi: 10.1371/journal.pone.0173766

Choi, I.-S., Jansen, R., and Ruhlman, T. (2019). Lost and found: return of the inverted repeat in the legume clade defined by its absence. *Genome Biol. Evol.* 11, 1321–1333. doi: 10.1093/gbe/evz076

Choi, I.-S., Jansen, R., and Ruhlman, T. (2020a). Caught in the act: variation in plastid genome inverted repeat expansion within and between populations of *Medicago minima*. *Ecol. Evol.* 10, 12129–12137. doi: 10.1002/ece3.6839

Choi, I.-S., Ruhlman, T. A., and Jansen, R. K. (2020b). Comparative mitogenome analysis of the genus *Trifolium* reveals independent gene fission of *ccmFn* and intracellular gene transfers in Fabaceae. *Int. J. Mol. Sci.* 21:1959. doi: 10.3390/ijms21061959

Choi, I.-S., Wojciechowski, M. F., Ruhlman, T. A., and Jansen, R. K. (2021). In and out: evolution of viral sequences in the mitochondrial genomes of legumes (Fabaceae). *Mol. Phylogenet. Evol.* 163:107236. doi: 10.1016/j.ympev.2021.107236

Chomicki, G., Ward, P. S., and Renner, S. S. (2015). Macroevolutionary assembly of ant/plant symbioses: *Pseudomyrmex* ants and their ant-housing plants in the Neotropics. *Proc. Royal Soc. B* 282:20152200. doi: 10.1098/rspb.2015.2200

Corriveau, J. L., and Coleman, A. W. (1988). Rapid screening method to detect potential biparental inheritance of plastid DNA and results for over 200 angiosperm species. *Am. J. Bot.* 75, 1443–1458. doi: 10.1002/j.1537-2197.1988.tb11219.x

Cowan, R. S. (1981). "Swartzieae," in *Advances in Legume Systematics, Part 1*, eds R. M. Polhill and P. H. Raven (Richmond, UK: Royal Botanic Gardens, Kew), 209–212.

Darling, A. E., Mau, B., and Perna, N. T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* 5:e11147. doi: 10.1371/journal.pone.0011147

Davis, C. C., Schaefer, H., Xi, Z., Baum, D. A., Donoghue, M. J., and Harmon, L. J. (2014). Long-term morphological stasis maintained by a plant-pollinator mutualism. *Proc. Natl. Acad. Sci. U. S. A.* 111, 5914–5919. doi: 10.1073/pnas.1403157111

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

Doyle, J. J., Chappill, J. A., Bailey, C. D., and Kajita, T. (2000). "Towards a comprehensive phylogeny of legumes: evidence from *rbcL* sequences and non-molecular data," in *Advances in Legume Systematics, Part 9*, eds P. S. Herendeen and A. Bruneau (Richmond, UK: Royal Botanic Gardens, Kew), 1–20.

Doyle, J. J., Davis, J. I., Soreng, R. J., Garvin, D., and Anderson, M. J. (1992). Chloroplast DNA inversions and the origin of the grass family (Poaceae). *Proc. Natl. Acad. Sci. U.S.A.* 89, 7722–7726. doi: 10.1073/pnas.89.16.7722

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

Doyle, J. J., Doyle, J. L., Ballenger, J. A., Dickson, E. E., Kajita, T., and Ohashi, H. (1997). A phylogeny of the chloroplast gene *rbcL* in the Leguminosae: taxonomic correlations and insights into the evolution of nodulation. *Am. J. Bot.* 84, 541–554. doi: 10.2307/2446030

Doyle, J. J., Doyle, J. L., Ballenger, J. A., and Palmer, J. D. (1996). The distribution and phylogenetic significance of a 50-kb chloroplast DNA inversion in the flowering plant family Leguminosae. *Mol. Phylogenet. Evol.* 5, 429–438. doi: 10.1006/mpev.1996.0038

Doyle, J. J., Doyle, J. L., and Palmer, J. D. (1995). Multiple independent losses of two genes and one intron from legume chloroplast genomes. *Syst. Bot.* 20, 272–294. doi: 10.2307/2419496

DRYFLOR (2016). Plant diversity patterns in neotropical dry forests and their conservation implications. *Science* 353, 1383–1387. doi: 10.1126/science.aaf5080

Duan, L., Harris, A., Ye, W., Deng, S.-W., Song, Z.-Q., Chen, H.-F., et al. (2019). Untangling the taxonomy of the Cladrastis clade (Leguminosae: papilionoideae) by integrating phylogenetics and ecological evidence. *Taxon* 68, 1189–1203. doi: 10.1002/tax.12155

Duan, L., Li, S.-J., Su, C., Sirichamorn, Y., Han, L.-N., Ye, W., et al. (2021). Phylogenomic framework of the IRLC legumes (Leguminosae subfamily Papilionoideae) and intercontinental biogeography of tribe Wisterieae. *Mol. Phylogenet. Evol.* 163:107235. doi: 10.1016/j.ympev.2021.107235

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

Egan, A. N., and Pan, B. (2015). Resolution of polyphyly in *Pueraria* (Leguminosae, Papilionoideae): the creation of two new genera, *Haymondia* and *Toxicopueraria*, the resurrection of *Neustanthus*, and a new combination in *Teyleria*. *Phytotaxa* 218, 201–226. doi: 10.11646/PHYTOTAXA.218.3.1

Egan, A. N., Vatanparast, M., and Cagle, W. (2016). Parsing polyphyletic *Pueraria*: delimiting distinct evolutionary lineages through phylogeny. *Mol. Phylogenet. Evol.* 104, 44–59. doi: 10.1016/j.ympev.2016.08.001

Ellison, N. W., Liston, A., Steiner, J. J., Williams, W. M., and Taylor, N. L. (2006). Molecular phylogenetics of the clover genus (*Trifolium*—Leguminosae). *Mol. Phylogenet. Evol.* 39, 688–705. doi: 10.1016/j.ympev.2006.01.004

Endress, P. K. (1996). *Diversity and Evolutionary Biology of Tropical Flowers*. Cambridge, UK: Cambridge University Press.

Endress, P. K. (1999). Symmetry in flowers: diversity and evolution. *Int. J. Plant Sci.* 160, S3–S23. doi: 10.1086/314211

Gepts, P., Beavis, W. D., Brummer, E. C., Shoemaker, R. C., Stalker, H. T., Weeden, N. F., et al. (2005). Legumes as a model plant family. Genomics for food and feed report of the cross-legume advances through genomics conference. *Plant Physiol.* 137, 1228–1235. doi: 10.1104/pp.105.060871

Gonçalves, D. J. P., Jansen, R. K., Ruhlman, T. A., and Mandel, J. R. (2020). Under the rug: abandoning persistent misconceptions that obfuscate organelle evolution. *Mol. Phylogenet. Evol.* 151:106903. doi: 10.1016/j.ympev.2020.106903

Gonçalves, D. J. P., Simpson, B. B., Ortiz, E. M., Shimizu, G. H., and Jansen, R. K. (2019). Incongruence between gene trees and species trees and phylogenetic signal variation in plastid genes. *Mol. Phylogenet. Evol.* 138, 219–232. doi: 10.1016/j.ympev.2019.05.022

Goremykin, V. V., Nikiforova, S. V., Cavalieri, D., Pindo, M., and Lockhart, P. (2015). The root of flowering plants and total evidence. *Syst. Biol.* 64, 879–891. doi: 10.1093/sysbio/syv028

Hazle, T., and Bonen, L. (2007). Status of genes encoding the mitochondrial S1 ribosomal protein in closely-related legumes. *Gene* 405, 108–116. doi: 10.1016/j.gene.2007.09.019

Hu, J.-M., Lavin, M., Wojciechowski, M. F., and Sanderson, M. J. (2000). Phylogenetic systematics of the tribe Millettieae (Leguminosae) based on chloroplast *trnK*/*matK* sequences and its implications for evolutionary patterns in Papilionoideae. *Am. J. Bot.* 87, 418–430. doi: 10.2307/2656638

Ireland, H., Pennington, R. T., and Preston, J. (2000). "Molecular systematics of the Swartzieae," in *Advances in Legume Systematics, Part 9*, eds P. S. Herendeen and A. Bruneau (Richmond, UK: Royal Botanic Gardens, Kew), 217–231.

Jansen, R. K., and Palmer, J. D. (1987). A chloroplast DNA inversion marks an ancient evolutionary split in the sunflower family (Asteraceae). *Proc. Natl. Acad. Sci. U.S.A.* 84, 5818–5822. doi: 10.1073/pnas.84.16.5818

Jansen, R. K., Wojciechowski, M. F., Sanniyasi, E., Lee, S.-B., and Daniell, H. (2008). Complete plastid genome sequence of the chickpea *Cicer arietinum* and the phylogenetic distribution of *rps12* and *clpP* intron losses among legumes (Leguminosae). *Mol. Phylogenet. Evol.* 48, 1204–1217. doi: 10.1016/j.ympev.2008.06.013

Janzen, D. H. (1966). Coevolution of mutualism between ants and acacias in Central America. *Evolution* 20, 249–275. doi: 10.2307/2406628

Jin, D.-P., Choi, I.-S., and Choi, B.-H. (2019). Plastid genome evolution in tribe Desmodieae (Fabaceae: papilionoideae). *PLoS One* 14:e0218743. doi: 10.1371/journal.pone.0218743

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

Kajita, T., Ohashi, H., Tateishi, Y., Bailey, C. D., and Doyle, J. J. (2001). *rbcL* and legume phylogeny, with particular reference to Phaseoleae, Millettieae, and allies. *Syst. Bot.* 26, 515–536. doi: 10.1043/0363-6445-26.3.515

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

Kite, G. C., and Pennington, R. T. (2003). Quinolizidine alkaloid status of *Styphnolobium* and *Cladrastis* (Leguminosae). *Biochem. Syst. Ecol.* 31, 1409–1416. doi: 10.1016/S0305-1978(03)00118-2

Kite, G. C., Veitch, N. C., Soto-Hernández, M., and Lewis, G. P. (2013). Highly glycosylated flavonols at the genistoid boundary and the systematic position of *Dermatophyllum*. *S. Afr. J. Bot.* 89, 181–187. doi: 10.1016/j.sajb.2013.06.003

Klitgård, B. B., Forest, F., Booth, T. J., and Saslis-Lagoudakis, C. H. (2013). A detailed investigation of the Pterocarpus clade (Leguminosae: dalbergieae): *etaballia* with radially symmetrical flowers is nested within the papilionoid-flowered *Pterocarpus*. *S. Afr. J. Bot.* 89, 128–142. doi: 10.1016/j.sajb.2013.07.006

Koenen, E. J. M., de Vos, J. M., Atchison, G. W., Simon, M. F., Schrire, B. D., de Souza, E. R., et al. (2013). Exploring the tempo of species diversification in legumes. *S. Afr. J. Bot.* 89, 19–30. doi: 10.1016/j.sajb.2013.07.005

Koenen, E. J. M., Kidner, C., Souza, E. R., Simon, M. F., Iganci, J. R., Nicholls, J. A., et al. (2020a). Hybrid capture of 964 nuclear genes resolves evolutionary relationships in the mimosoid legumes and reveals the polytomous origins of a large pantropical radiation. *Am. J. Bot.* 107, 1710–1735. doi: 10.1002/ajb2.1568

Koenen, E. J. M., Ojeda, D. I., Steeves, R., Migliore, J., Bakker, F. T., Wieringa, J. J., et al. (2020b). 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

Koenen, E. J. M., Ojeda, D. I., Bakker, F. T., Wieringa, J. J., Kidner, C., Hardy, O. J., et al. (2021). The origin of the legumes is a complex paleopolyploid phylogenomic tangle closely associated with the Cretaceous–Paleogene (K–Pg) mass extinction event. *Syst. Biol.* 70, 508–526. doi: 10.1093/sysbio/syaa041

Kursar, T. A., Dexter, K. G., Lokvam, J., Pennington, R. T., Richardson, J. E., Weber, M. G., et al. (2009). The evolution of antiherbivore defenses and their contribution to species coexistence in the tropical tree genus *Inga*. *Proc. Natl. Acad. Sci. U.S.A.* 106, 18073–18078. doi: 10.1073/pnas.0904786106

Lackey, J. A. (1981). "Phaseoleae," in *Advances in Legume Systematics, Part 1*, eds R. M. Polhill and P. H. Raven (Richmond, UK: Royal Botanic Gardens, Kew), 301–327.

Lai, M., Sceppa, J., Ballenger, J. A., Doyle, J. J., and Wunderlin, R. P. (1997). Polymorphism for the presence of the *rpl2* intron in chloroplast genomes of *Bauhinia* (Leguminosae). *Syst. Bot.* 22, 519–528. doi: 10.2307/2419825

Lavin, M., Doyle, J. J., and Palmer, J. D. (1990). Evolutionary significance of the loss of the chloroplast-DNA inverted repeat in the Leguminosae subfamily Papilionoideae. *Evolution* 44, 390–402. doi: 10.1111/j.1558-5646.1990.tb05207.x

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

Lavin, M., Pennington, R. T., Klitgaard, B. B., Sprent, J. I., Lima, H. C., and Gasson, P. E. (2001). The dalbergioid legumes (Fabaceae): delimitation of a pantropical monophyletic clade. *Am. J. Bot.* 88, 503–533. doi: 10.2307/2657116

Lee, C., Choi, I.-S., Cardoso, D., Lima, H. C., Queiroz, L. P., Wojciechowski, M. F., et al. (2021). The chicken or the egg? Plastome evolution and an independent loss of the inverted repeat in papilionoid legumes. *Plant J.* 107, 861–875. doi: 10.1111/tpj.15351

Lee, S. T., Cook, D., Molyneux, R. J., Davis, T. Z., and Gardner, D. R. (2013). Alkaloid profiles of *Dermatophyllum arizonicum*, *Dermatophyllum gypsophilum*, *Dermatophyllum secundiflorum*, *Styphnolobium affine*, and *Styphnolobium japonicum* previously classified as *Sophora* species. *Biochem. Syst. Ecol.* 49, 87–93. doi: 10.1016/j.bse.2013.03.018

LPWG [Legume Phylogeny Working Group] (2013). Legume phylogeny and classification in the 21st century: Progress, prospects and lessons for other species-rich clades. *Taxon* 62, 217–248. doi: 10.12705/622.8

LPWG [Legume Phylogeny Working Group] (2017). A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny. *Taxon* 66, 44–77. doi: 10.12705/661.3

LPWG [Legume Phylogeny Working Group] (2021). *The World Checklist of Vascular Plants (WCVP): Fabaceae, vers. June 2021,* ed R. Govaerts. Available online at: http://sftp.kew.org/pub/data_collaborations/Fabaceae/DwCA/

Leite, V. G., Mansano, V. F., and Teixeira, S. P. (2014). Floral ontogeny in Dipterygeae (Fabaceae) reveals new insights into one of the earliest branching tribes in papilionoid legumes. *Bot. J. Linnean Soc.* 174, 529–550. doi: 10.1111/boj.12158

Leite, V. G., Teixeira, S. P., Mansano, V. F., and Prenner, G. (2015). Floral development of the early-branching papilionoid legume *Amburana cearensis* (Leguminosae) reveals rare and novel characters. *Int. J. Plant Sci.* 176, 94–106. doi: 10.1086/678468

Letunic, I., and Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* 49, W293–W296. doi: 10.1093/nar/gkab301

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

Li, H.-T., Luo, Y., Gan, L., Ma, P.-F., Gao, L.-M., Yang, J.-B., et al. (2021). Plastid phylogenomic insights into relationships of all flowering plant families. *BMC Biol.* 19:232. doi: 10.1186/s12915-021-01166-2

Lima, H. C. (1980). Revisão taxonômica do gênero *Vataireopsis* Ducke (Leguminosae–Faboideae). *Rodriguésia* 32, 21–40.

Lima, H. C. (1982). Revisão taxonômica do gênero *Vatairea* Aubl. (Leguminosae–Faboideae). *Arch. Jard. Bot. Rio de Janeiro* 26, 173–214.

Lima, H. C. (1990). Tribo Dalbergieae (Leguminosae Papilionoideae)–Morfologia dos frutos, sementes e plântulas e sua aplicação na sistemática. *Arch. Jard. Bot. Rio de Janeiro* 30, 1–42.

Marazzi, B., Ané, C., Simon, M. F., Delgado-Salinas, A., Luckow, M., and Sanderson, M. J. (2012). Locating evolutionary precursors on a phylogenetic tree. *Evolution* 66, 3918–3930. doi: 10.1111/j.1558-5646.2012.01720.x

Marazzi, B., Gonzalez, A. M., Delgado-Salinas, A., Luckow, M. A., Ringelberg, J. J., and Hughes, C. E. (2019). Extrafloral nectaries in Leguminosae: phylogenetic distribution, morphological diversity and evolution. *Aust. Syst. Bot.* 32, 409–458. doi: 10.1071/SB19012

Marazzi, B., and Sanderson, M. J. (2010). Large-scale patterns of diversification in the widespread legume genus *Senna* and the evolutionary role of extrafloral nectaries. *Evolution* 64, 3570–3592. doi: 10.1111/j.1558-5646.2010.01086.x

Martin, G. E., Rousseau-Gueutin, M., Cordonnier, S., Lima, O., Michon-Coudouel, S., Naquin, D., et al. (2014). The first complete chloroplast genome of the Genistoid legume *Lupinus luteus*: evidence for a novel major lineage-specific rearrangement and new insights regarding plastome evolution in the legume family. *Ann. Bot.* 113, 1197–1210. doi: 10.1093/aob/mcu050

McMahon, M., and Hufford, L. (2004). Phylogeny of Amorpheae (Fabaceae: papilionoideae). *Am. J. Bot.* 91, 1219–1230. doi: 10.3732/ajb.91.8.1219

McMahon, M. M., and Sanderson, M. J. (2006). Phylogenetic supermatrix analysis of GenBank sequences from 2228 papilionoid legumes. *Syst. Biol.* 55, 818–836. doi: 10.1080/10635150600999150

Milligan, B. G., Hampton, J. N., and Palmer, J. D. (1989). Dispersed repeats and structural reorganization in subclover chloroplast DNA. *Mol. Biol. Evol.* 6, 355–368. doi: 10.1093/oxfordjournals.molbev.a040558

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

Ojeda, D. I, Koenen, E. J. M., Cervantes, S., de la Estrella, M., Banguera-Hinestroza, E., Janssens, S. B., et al. (2019). Phylogenomic analyses reveal an exceptionally high number of evolutionary shifts in a florally diverse clade of African legumes. *Mol. Phylogenet. Evol.* 137, 156–167. doi: 10.1016/j.ympev.2019.05.002

Orton, L. M., Barberá, P., Nissenbaum, M. P., Peterson, P. M., Quintanar, A., Soreng, R. J., et al. (2021). A 313 plastome phylogenomic analysis of Pooideae: exploring relationships among the largest subfamily of grasses. *Mol. Phylogenet. Evol.* 159:107110. doi: 10.1016/j.ympev.2021.107110

Oyebanji, O., Zhang, R., Chen, S.-Y., and Yi, T.-S. (2020). New insights into the plastome evolution of the Millettioid/Phaseoloid clade (Papilionoideae, Leguminosae). *Front. Plant. Sci.* 11:151. doi: 10.3389/fpls.2020.00151

Palmer, J. D., Osorio, B., Aldrich, J., and Thompson, W. F. (1987). Chloroplast DNA evolution among legumes: loss of a large inverted repeat occurred prior to other sequence rearrangements. *Curr. Genet.* 11, 275–286. doi: 10.1007/BF00355401

Palmer, J. D., and Thompson, W. F. (1982). Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost. *Cell* 29, 537–550. doi: 10.1016/0092-8674(82)90170-2

Park, S., An, B., and Park, S. (2020). Recurrent gene duplication in the angiosperm tribe Delphinieae (Ranunculaceae) inferred from intracellular gene transfer events and heteroplasmic mutations in the plastid *matK* gene. *Sci. Rep.* 10:2720. doi: 10.1038/s41598-020-59547-6

Pennington, R. T., Klitgaard, B. B., Ireland, H., and Lavin, M. (2000). "New insights into floral evolution of basal Papilionoideae from molecular phylogenies," in *Advances in Legume Systematics, Part 9*, eds P. S. Herendeen and A. Bruneau (Richmond, UK: Royal Botanic Gardens, Kew), 233–248.

Pennington, R. T., Lavin, M., Ireland, H., Klitgaard, B., Preston, J., and Hu, J.-M. (2001). Phylogenetic relationships of basal papilionoid legumes based upon sequences of the chloroplast *trnL* intron. *Syst. Bot.* 26, 537–556. doi: 10.1043/0363-6445-26.3.537

Pereira, J. B. S., Giulietti, A. M., Prado, J., Vasconcelos, S., Watanabe, M. T. C., Pinangé, D. S. B., et al. (2021). Plastome-based phylogenomics elucidate relationships in rare *Isoëtes* species groups from the Neotropics. *Mol. Phylogenet. Evol.* 161:107177. doi: 10.1016/j.ympev.2021.107177

Polhill, R. M. (1981a). "Sophoreae," in *Advances in Legume Systematics, Part 1*, eds R. M. Polhill and P. H. Raven (Richmond, UK: Royal Botanic Gardens, Kew), 213–230.

Polhill, R. M. (1981b). "Papilionoideae," in *Advances in Legume Systematics, Part 1*, eds R. M. Polhill and P. H. Raven (Richmond, UK: Royal Botanic Gardens, Kew), 191–208.

Polhill, R. M. (1994). "Classification of the Leguminosae," in *Phytochemical Dictionary of the Leguminosae*, eds F. A. Bisby, J. Buckingham, and J. B. Harborne (London, UK: Chapman and Hall), xxxv–xlviii.

Prenner, G., Cardoso, D., Zartman, C. E., and Queiroz, L. P. (2015). Flowers of the early-branching papilionoid legume *Petaladenium urceoliferum* display unique morphological and ontogenetic features. *Am. J. Bot.* 102, 1780–1793. doi: 10.3732/ajb.1500348

Queiroz, L. P., Pastore, J. F. B., Cardoso, D., Snak, C., Lima, A. L. D. C., Gagnon, E., et al. (2015). A multilocus phylogenetic analysis reveals the monophyly of a recircumscribed papilionoid legume tribe Diocleae with well-supported generic

relationships. *Mol. Phylogenet. Evol.* 90, 1–19. doi: 10.1016/j.ympev.2015.04.016

Queiroz, L. P., São Mateus, W., Delgado-Salinas, A., Torke, B. M., Lewis, G. P., Dorado, O., et al. (2017). A molecular phylogeny reveals the Cuban enigmatic genus *Behaimia* as a new piece in the Brongniartieae puzzle of papilionoid legumes. *Mol. Phylogenet. Evol.* 109, 191–202. doi: 10.1016/j.ympev.2017.01.001

Ramos, G., Lima, H. C., Prenner, G., Queiroz, L. P., Zartman, C. E., and Cardoso, D. (2016). Molecular systematics of the Amazonian genus *Aldina*, a phylogenetically enigmatic ectomycorrhizal lineage of papilionoid legumes. *Mol. Phylogenet. Evol.* 97, 11–18. doi: 10.1016/j.ympev.2015.12.017

Regier, J. C., Shultz, J. W., Zwick, A., Hussey, A., Ball, B., Wetzer, R., et al. (2010). Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature* 463, 1079–1083. doi: 10.1038/nature08742

Sabir, J., Schwarz, E., Ellison, N., Zhang, J., Baeshen, N. A., Mutwakil, M., et al. (2014). Evolutionary and biotechnology implications of plastid genome variation in the inverted-repeat-lacking clade of legumes. *Plant Biotechnol. J.* 12, 743–754. doi: 10.1111/pbi.12179

Sanderson, M. J., and Donoghue, M. J. (1994). Shifts in diversification rate with the origin of angiosperms. *Science* 264, 1590–1593. doi: 10.1126/science.264.5165.1590

Sargent, R. D. (2004). Floral symmetry affects speciation rates in angiosperms. *Proc. R. Soc. Lond. B* 271, 603–608. doi: 10.1098/rspb.2003.2644

Schaefer, H., Hechenleitner, P., Santos-Guerra, A., Sequeira, M. M., Pennington, R. T., Kenicer, G., et al. (2012). Systematics, biogeography, and character evolution of the legume tribe Fabeae with special focus on the middle-Atlantic island lineages. *BMC Evol. Biol.* 12:250. doi: 10.1186/1471-2148-12-250

Schneider, J. V., Paule, J., Jungcurt, T., Cardoso, D., Amorim, A. M., Berberich, T., et al. (2021). Resolving recalcitrant clades in the Pantropical Ochnaceae: insights from comparative phylogenomics of plastome and nuclear genomic data derived from targeted sequencing. *Front. Plant Sci.* 12:638650. doi: 10.3389/fpls.2021.638650

Schrire, B. D., Lavin, M., Barker, N. P., and Forest, F. (2009). Phylogeny of the tribe Indigofereae (Leguminosae–Papilionoideae): geographically structured more in succulent-rich and temperate settings than in grass-rich environments. *Am. J. Bot.* 96, 816–852. doi: 10.3732/ajb.0800185

Schrire, B. D., Lavin, M., and Lewis, G. P. (2005). Global distribution patterns of the Leguminosae: insights from recent phylogenies. *Biol. Skr.* 55, 375–422.

Schwarz, E. N., Ruhlman, T. A., Sabir, J. S., Hajrah, N. H., Alharbi, N. S., Al-Malki, A. L., et al. (2015). Plastid genome sequences of legumes reveal parallel inversions and multiple losses of *rps16* in papilionoids. *J. Syst. Evol.* 53, 458–468. doi: 10.1111/jse.12179

Serna-Sánchez, M. A., Pérez-Escobar, O. A., Bogarín, D., Torres-Jimenez, M. F., Alvarez-Yela, A. C., Arcila-Galvis, J. E., et al. (2021). Plastid phylogenomics resolves ambiguous relationships within the orchid family and provides a solid timeframe for biogeography and macroevolution. *Sci. Rep.* 11:6858. doi: 10.1038/s41598-021-83664-5

Sinjushin, A. A. (2018). Floral ontogeny in *Cordyla pinnata* (A. Rich.) Milne-Redh. (Leguminosae, Papilionoideae): away from stability. *Flora* 241, 8–15. doi: 10.1016/j.flora.2018.02.005

Specht, C. D., and Bartlett, M. E. (2009). Flower evolution: the origin and subsequent diversification of the angiosperm flower. *Annu. Rev. Ecol. Evol. Syst.* 40, 217–243. doi: 10.1146/annurev.ecolsys.110308.120203

Sprent, J. I. (2000). "Nodulation as a taxonomic tool," in *Advances in Legume Systematics, Part 9*, eds P. S. Herendeen and A. Bruneau (Richmond, UK: Royal Botanic Gardens, Kew), 21–43.

Sprent, J. I., Ardley, J., and James, E. K. (2017). Biogeography of nodulated legumes and their nitrogen-fixing symbionts. *New Phytol.* 215, 40–56. doi: 10.1111/nph.14474

Steele, K. P., and Wojciechowski, M. F. (2003). "Phylogenetic analyses of tribes Trifolieae and Vicieae, based on sequences of the plastid gene *matK* (Papilionoideae: leguminosae)," in *Advances in Legume Systematics, Part 10*, eds B. B. Klitgaard and A. Bruneau (Richmond, UK: Royal Botanic Gardens, Kew), 355–370.

Stefanović, S., Pfeil, B. E., Palmer, J. D., and Doyle, J. J. (2009). Relationships among phaseoloid legumes based on sequences from eight chloroplast regions. *Syst. Bot.* 34, 115–128. doi: 10.1600/036364409787602221

Sveinsson, S., and Cronk, Q. (2014). Evolutionary origin of highly repetitive plastid genomes within the clover genus (*Trifolium*). *BMC Evol. Biol.* 14:228. doi: 10.1186/s12862-014-0228-6

Swanepoel, W., Le Roux, M. M., Wojciechowski, M. F., and Van Wyk, A. E. (2015). *Oberholzeria* (Fabaceae subfam. Faboideae), a new monotypic legume genus from Namibia. *PLoS One* 10:e0122080. doi: 10.1371/journal.pone.0122080

Thode, V. A., Lohmann, L. G., and Sanmartín, I. (2020). Evaluating character partitioning and molecular models in plastid phylogenomics at low taxonomic levels: a case study using *Amphilophium* (Bignonieae, Bignoniaceae). *J. Syst. Evol.* 58, 1071–1089. doi: 10.1111/jse.12579

Tucker, S. C. (1993). Floral ontogeny in Sophoreae (Leguminosae: papilionoideae). I. *Myroxylon* (Myroxylon group) and *Castanospermum* (Angylocalyx group). *Am. J. Bot.* 80, 65–75. doi: 10.1002/j.1537-2197.1993.tb13768.x

Tucker, S. C. (2003a). Floral development in legumes. *Plant Physiol.* 131, 911–926. doi: 10.1104/pp.102.017459

Tucker, S. C. (2003b). Floral ontogeny in *Swartzia* (Leguminosae: papilionoideae: swartzieae): distribution and role of the ring meristem. *Am. J. Bot.* 90, 1271–1292. doi: 10.3732/ajb.90.9.1271

Tucker, S. C., and Douglas, A. W. (1994). "Ontogenetic evidence and phylogenetic relationships among basal taxa of legumes," in *Advances in Legume Systematics, Part 6*, eds I. K. Ferguson and S. C. Tucker (Richmond, UK: Royal Botanic Gardens, Kew), 11–32.

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

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

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

Wink, M. (2013). Evolution of secondary metabolites in legumes (Fabaceae). *S. Afr. J. Bot.* 89, 164–175. doi: 10.1016/j.sajb.2013.06.006

Wink, M., Botschen, F., Gosmann, C., Schäfer, H., and Waterman, P. G. (2010). "Chemotaxonomy seen from a phylogenetic perspective and evolution of secondary metabolism," in *Annual Plant Reviews Volume 40: Biochemistry of Plant Secondary Metabolism*, ed. M. Wink (Oxford, UK: Wiley-Blackwell), 364–433.

Wojciechowski, M. F. (2013a). The origin and phylogenetic relationships of the Californian chaparral 'paleoendemic' *Pickeringia* (Leguminosae). *Syst. Bot.* 38, 132–142. doi: 10.1600/036364413X662024

Wojciechowski, M. F. (2013b). Towards a new classification of Leguminosae: naming clades using non-Linnaean phylogenetic nomenclature. *S. Afr. J. Bot.* 89, 85–93. doi: 10.1016/j.sajb.2013.06.017

Wojciechowski, M. F., Lavin, M., and Sanderson, M. J. (2004). A phylogeny of legumes (Leguminosae) based on analysis of the plastid *matK* gene resolves many well-supported subclades within the family. *Am. J. Bot.* 91, 1846–1862. doi: 10.3732/ajb.91.11.1846

Wojciechowski, M. F., Sanderson, M. J., Steele, K. P., and Liston, A. (2000). "Molecular phylogeny of the "temperate herbaceous tribes" of papilionoid legumes: a supertree approach," in *Advances in Legume Systematics, Part 9*, eds P. S. Herendeen and A. Bruneau (Richmond, UK: Royal Botanic Gardens, Kew), 277–298.

Xi, Z., Ruhfel, B. R., Schaefer, H., Amorim, A. M., Sugumaran, M., Wurdack, K. J., et al. (2012). Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. *Proc. Natl. Acad. Sci. U.S.A.* 109, 17519–17524. doi: 10.1073/pnas.1205818109

Yahara, T., Javadi, F., Onoda, Y., Queiroz, L. P., Faith, D. P., Prado, D. E., et al. (2013). Global legume diversity assessment: concepts, key indicators, and strategies. *Taxon* 62, 249–266. doi: 10.12705/622.12

Yang, L., Su, D., Chang, X., Foster, C. S. P., Sun, L., Huang, C.-H., et al. (2020). Phylogenomic insights into deep phylogeny of angiosperms based on broad nuclear gene sampling. *Plant Commun.* 1:100027. doi: 10.1016/j.xplc.2020.100027

Yu, G., Smith, D., Zhu, H., Guan, Y., and Lam, T. T.-Y. (2017). ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Meth. Ecol. Evol.* 8, 28–36. doi: 10.1111/2041-210X.12628

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

Zhang, Q., Liu, Y., and Sodmergen (2003). Examination of the cytoplasmic DNA in male reproductive cells to determine the potential for cytoplasmic inheritance in 295 angiosperm species. *Plant Cell Physiol.* 44, 941–951. doi: 10.1093/pcp/pcg121

Zhang, R., Wang, Y.-H., Jin, J.-J., Stull, G. W., Bruneau, A., Cardoso, D., et al. (2020). Exploration of plastid phylogenomic conflict yields new insights into the deep relationships of Leguminosae. *Syst. Biol.* 69, 613–622.

Zhao, Y., Zhang, R., Jiang, K.-W., Qi, J., Hu, Y., Guo, J., et al. (2021). Nuclear phylotranscriptomics and phylogenomics support numerous polyploidization events and hypotheses for the evolution of rhizobial nitrogen-fixing symbiosis in Fabaceae. *Mol. Plant.* 14, 748–773. doi: 10.1016/j.molp.2021.02.006

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

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

Copyright © 2022 Choi, Cardoso, de Queiroz, de Lima, Lee, Ruhlman, Jansen and Wojciechowski. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Plant Science

ERRATUM
published: 24 May 2022
doi: 10.3389/fpls.2022.930260

Erratum: Highly Resolved Papilionoid Legume Phylogeny Based on Plastid Phylogenomics

*Frontiers Production Office**

Frontiers Media SA, Lausanne, Switzerland

Keywords: deep evolution, Meso-Papilionoideae, plastid genome, Papilionoideae, Leguminosae

An Erratum on

Highly Resolved Papilionoid Legume Phylogeny Based on Plastid Phylogenomics
by Choi, I.-S., Cardoso, D., de Queiroz, L. P., de Lima, H. C., Lee, C., Ruhlman, T. A., Jansen, R. K., and Wojciechowski, M. F. (2022). Front. Plant Sci. 13:823190. doi: 10.3389/fpls.2022.823190

Due to a production error, the reference and citation for "DRYFLOR, 2016" was incorrectly written as "DRYFLOR et al., 2016" and "Dryflor, Banda-R. K., Delgado-Salinas, A., Dexter, K. G., Linares-Palomino, R., and Oliveira-Filho, A. (2016). Plant diversity patterns in neotropical dry forests and their conservation implications. *Science* 353, 1383–1387. doi: 10.1126/science.aaf508". It should be "DRYFLOR, 2016" and "DRYFLOR (2016). Plant diversity patterns in neotropical dry forests and their conservation implications. *Science* 353, 1383–1387. doi: 10.1126/science.aaf5080".

Due to a production error, the reference and citation for "LPWG, 2013" was incorrectly written as "LPWG et al., 2013" and "LPWG, Bruneau, A., Doyle, J. J., Herendeen, P., Hughes, C., and Kenicer, G. (2013). Legume phylogeny and classification in the 21st century: progress, prospects and lessons for other species-rich clades. *Taxon* 62, 217–248. doi: 10.12705/622.8". It should be "LPWG, 2013" and "LPWG [Legume Phylogeny Working Group] (2013). Legume phylogeny and classification in the 21st century: Progress, prospects and lessons for other species-rich clades. *Taxon* 62, 217–248. doi: 10.12705/622.8".

Due to a production error, the reference and citation for "LPWG, 2017" was incorrectly written as "LPWG et al., 2017" and LPWG, Azani, N., Babineau, M., Bailey, C. D., Banks, H., Barbosa, A. R., et al. (2017). A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny. *Taxon* 66, 44–77. doi: 10.12705/661.3". It should be "LPWG, 2017" and "LPWG [Legume Phylogeny Working Group] (2017). A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny. *Taxon* 66, 44–77. doi: 10.12705/661.3".

Due to a production error, the reference for "LPWG, 2021" was incorrectly written as "LPWG, Bruneau, A., Doyle, J. J., Herendeen, P., Hughes, C., and Kenicer, G. (2013). Legume phylogeny and classification in the 21st century: progress, prospects and lessons for other species-rich clades. Taxon 62, 217–248. doi: 10.12705/622.8". It should be "LPWG, 2021. *The World Checklist of Vascular Plants (WCVP): Fabaceae, vers. June 2021*, ed R. Govaerts. Available online at: http://sftp.kew.org/pub/data_collaborations/Fabaceae/DwCA/."

The publisher apologizes for these mistakes. The original version of this article has been updated.

OPEN ACCESS

Approved by:
Frontiers Editorial Office, Frontiers Media SA, Switzerland

***Correspondence:**
Frontiers Production Office
production.office@frontiersin.org

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

Received: *27 April 2022*
Accepted: *27 April 2022*
Published: *24 May 2022*

Citation:
Frontiers Production Office (2022) Erratum: Highly Resolved Papilionoid Legume Phylogeny Based on Plastid Phylogenomics. Front. Plant Sci. 13:930260. doi: 10.3389/fpls.2022.930260

REFERENCES

DRYFLOR (2016). Plant diversity patterns in neotropical dry forests and their conservation implications. *Science* 353, 1383–1387. doi: 10.1126/science.aaf5080

LPWG [Legume Phylogeny Working Group] (2013). Legume phylogeny and classification in the 21st century: Progress, prospects and lessons for other species-rich clades. *Taxon* 62, 217–248. doi: 10.12705/622.8

LPWG [Legume Phylogeny Working Group] (2017). A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny. *Taxon* 66, 44–77. doi: 10.12705/661.3

LPWG [Legume Phylogeny Working Group] (2021). *The World Checklist of Vascular Plants (WCVP): Fabaceae, vers. June 2021*, ed R. Govaerts. Available online at: http://sftp.kew.org/pub/data_collaborations/Fabaceae/DwCA/

Copyright © 2022 Frontiers Production Office. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 15 June 2022
doi: 10.3389/fpls.2022.898881

Karyology and Genome Size Analyses of Iranian Endemic *Pimpinella* (Apiaceae) Species

*Shaghayegh Mehravi[1,2], Gholam Ali Ranjbar[2], Hamid Najafi-Zarrini[2], Ghader Mirzaghaderi[3], Mehrdad Hanifei[4], Anita Alice Severn-Ellis[1], David Edwards[1] and Jacqueline Batley[1]**

[1] School of Biological Sciences, University of Western Australia, Perth, WA, Australia, [2] Department of Plant Breeding and Biotechnology, Faculty of Crop Sciences, Sari Agricultural Sciences and Natural Resources University, Sari, Iran, [3] Department of Agronomy and Plant Breeding, Faculty of Agriculture, University of Kurdistan, Kurdistan, Iran, [4] Department of Plant Genetics and Breeding, Faculty of Agriculture, Tarbiat Modares University, Tehran, Iran

Pimpinella species are annual, biennial, and perennial semibushy aromatic plants cultivated for folk medicine, pharmaceuticals, food, and spices. The karyology and genome size of 17 populations of 16 different *Pimpinella* species collected from different locations in Iran were analyzed for inter-specific karyotypic and genome size variations. For karyological studies, root tips were squashed and painted with a DAPI solution (1 mg/ml). For flow cytometric measurements, fresh leaves of the standard reference (*Solanum lycopersicum* cv. Stupick, 2C DNA = 1.96 pg) and the *Pimpinella* samples were stained with propidium iodide. We identified two ploidy levels: diploid (2*x*) and tetraploid (4*x*), as well as five metaphase chromosomal counts of 18, 20, 22, 24, and 40. 2*n* = 24 is reported for the first time in the *Pimpinella* genus, and the presence of a B-chromosome is reported for one species. The nuclear DNA content ranged from 2C = 2.48 to 2C = 5.50 pg, along with a wide range of genome sizes between 1212.72 and 2689.50 Mbp. The average monoploid genome size and the average value of 2C DNA/chromosome were not proportional to ploidy. There were considerable positive correlations between 2C DNA and total chromatin length and total chromosomal volume. The present study results enable us to classify the genus *Pimpinella* with a high degree of morphological variation in Iran. In addition, cytological studies demonstrate karyotypic differences between *P. anthriscoides* and other species of *Pimpinella*, which may be utilized as a novel identification key to affiliate into a distinct, new genus – *Pseudopimpinella*.

Keywords: ***Pimpinella*, chromosome, karyology, DNA *C*-value, genome size, B-chromosome**

OPEN ACCESS

Edited by:
Wellington Ronildo Clarindo, Universidade Federal de Viçosa, Brazil

Reviewed by:
Maria Andréia Corrêa Mendonça, Goiano Federal Institute (IFGOIANO), Brazil
Milene Miranda Praça-Fontes, Universidade Federal do Espírito Santo, Brazil
Mariana Cansian Sattler, Universidade Federal de Viçosa, Brazil

***Correspondence:**
Jacqueline Batley
jacqueline.batley@uwa.edu.au

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

Received: *18 March 2022*
Accepted: *05 May 2022*
Published: *15 June 2022*

Citation:
Mehravi S, Ranjbar GA, Najafi-Zarrini H, Mirzaghaderi G, Hanifei M, Severn-Ellis AA, Edwards D and Batley J (2022) Karyology and Genome Size Analyses of Iranian Endemic Pimpinella (Apiaceae) Species. Front. Plant Sci. 13:898881. doi: 10.3389/fpls.2022.898881

INTRODUCTION

The genus *Pimpinella* is one of the largest genera in the family Apiaceae, subfamily Apioideae, and tribe Pimpinelleae, with approximately 170–180 species distributed throughout Europe, Asia, Africa, and South and North America (Pimenov and Leonov, 1993; Pu and Watson, 2005). The members of this genus are annual, biennial or perennial, semibushy, and aromatic plants with cordate-ovoid or oblong-ovoid fruits with five filiform ribs on each cordate-ovoid or oblong-ovoid leaves (Pu and Watson, 2005). Nearly 23 species of this genus are grown in Iran, six of

which are endemic (*P. pastinacifolia*, *P. tragioides*, *P. khorasanica*, *P. deverroides*, *P. khayyamii,* and *P. anisactis*) (Mozaffarian, 2003). Their habitats are dry slopes, rocky crevices, fields, meadows, mountain pastures, grasslands, steppes, and dry open woodlands 1,000–2,200 m above sea level.

Studies suggest that the genus *Pimpinella* is highly diverse, and the taxonomic delimitation of the genus has not yet been resolved (Zhou et al., 2008, 2009; Downie et al., 2010). The last revision of the genus was made by Wolff (1927) based on the petal color, fruit and petal vestiture, and life history. The genus was subdivided into three sections: *Reutera*, *Tragium,* and *Tragoselinum*. It has since been realized that morphological markers do not explain the systematic relationship among *Pimpinella* species (Fereidounfar et al., 2016). Hence, investigation of various aspects, including karyological observations and genome size estimates, may be useful in establishing systematic and evolutionary relationships, resolving taxonomic ambiguities, and gaining a better understanding of the way they diverged from each other (Dobigny et al., 2004; Knight et al., 2005; Bancheva and Greilhuber, 2006; Guerra, 2008; Bainard et al., 2013).

Basic chromosome number (x) variations of *Pimpinella* species have been shown by previous karyotypic studies as follows: $x = 8$ for *P. affinis* (Yurtseva, 1988), $x = 9$ for *P. tragium*, *P. saxifrage* and *P. puberula* (Gawlowska, 1967; Yurtseva, 1988; Pimenov et al., 2003), $x = 10$ for *P. corymbosa* and *P. lutea* (Al-Eisawi, 1989; Verlaque and Filosa, 1992; Pimenov et al., 1996), $x = 11$ for *P. buchananii*, *P. trifurcate,* and *P. tragium* (Constance and Chuang, 1982; Abebe, 1992; Yurtseva and Tikhomirov, 1998; Pimenov et al., 2003) and $x = 18$ and 20 for *P. saxifrage* (Gawlowska, 1967). However, a literature review suggests that $x = 9$ is the plesiomorphic state within the *Pimpinella,* which – with subsequent aneuploidy – resulted in an increase to $x = 10$, 11, and 12 in the more derived species, which seems to be a plausible hypothesis. For *Pimpinella*, currently, the most accepted basic chromosome number is $x = 9$. Polyploidy and aneuploidy levels have also been reported for *Pimpinella* species (Constance and Chuang, 1982; Daushkevich et al., 1995; Shner et al., 2004). Overall, the extraordinary variations in *Pimpinella* chromosome number can be reflected in inter/intraspecific nuclear DNA contents. Nuclear DNA amounts have been estimated for *P. cumbrae* and *P. saxifrage* (2C DNA = 4.60 and 8.52 pg, respectively) by flow cytometry (FCM) (Grime and Mowforth, 1982; Suda et al., 2003; Temsch et al., 2010).

Nuclear DNA content is under strict genotypic control within the defined limits. Thus, it appears that such a variation correlates with evolutionary and systematic considerations (Bennett and Leitch, 2000; Greilhuber et al., 2005; Knight et al., 2005; Doležel et al., 2007; Bainard et al., 2013). Variation of intra/inter-specific genome size may reflect karyotypic differences, such as differences in the case of chromosome number and size (Bennett et al., 2008). Greilhuber et al. (2005) the DNA content of the unreplicated haploid chromosomal supplement, n, (1 C-value), and the amount of DNA per basic chromosome number, x, (1 Cx-value, regardless of generative polyploidy, aneuploidies, or other factors). Variation in chromosome number in the *Pimpinella* genus can indicate intra- and inter-specific differences in genomic DNA quantities.

No detailed information is available regarding the DNA C-value, karyology, and ploidy levels of *Pimpinella* species, as cytological investigations have mainly concentrated on reporting chromosome numbers. Therefore, this research reports for the first time the karyotype criteria and genome size of 16 Iranian species of *Pimpinella*.

MATERIALS AND METHODS

Plant Materials

The seeds of 16 Iranian endemic species of *Pimpinella* were collected during the growing season in their natural habitats from different locations in Iran. Only S8 was collected from two geographical locations. The species code and geographical descriptions, including latitude, longitude, altitude (m), mean temperature (°C), and mean rainfall (mm), are shown in **Table 1** and **Figure 1**.

Cytogenetic Analysis

Actively growing roots of approximately 1–2 cm were cut and pretreated in a 0.002 M solution of 8-hydroxyquinoline for 4 h at 25°C and fixed in ethanol: glacial acetic acid (3:1, v/v) for 24–36 h at 4°C. Chromosome preparations from root tip cells were performed, as described by Abdolmalaki et al. (2019). For each species, ten root tips were flooded with ice-cold water twice (5 min each time), followed by 0.01 M citrate buffer twice for 5 min. Between 1 and 1.5 mm of the root tips (meristematic parts) were then digested with a 30 μl enzyme mixture containing 1% pectolyase (Sigma P3026), 0.7% cellulose (CalBiochem219466), 0.7% cellulose R10 (Duchefa C8001), and 1% cytohelicase (Sigma C8274) dissolved in 0.01 M citrate buffer with pH 4.8 for 1 h. After digestion, meristems were washed twice with citrate buffer (5 min each time) and once with ethanol for 5 min to remove the enzyme mixture. Ethanol was changed with a 70 μl fixative solution (9: glacial acetic acid/1: absolute methanol). The root tips were carefully taped using a dissecting needle until a cell suspension formed. Seven microliters of the cell suspension were then dropped onto each glass slide in a box lined with 50% humidity; it was left to dry slowly and stored in 70% ethanol. A drop containing 1 μg/ml DAPI (4′, 6-diamidino-2-phenylindole) was added to the cell area, and a coverslip was applied. High-resolution chromosome images were taken using a Nikon A1Si Laser Scanning Confocal Microscope (Nikon Instruments Inc., Japan). Chromosome measurements and karyotypic features were studied based on five well-prepared metaphase plates from different individuals.

Karyotype Characterization

Chromosomal parameters were determined in each metaphase plate to assess the karyotypes numerically. These parameters included long arm (L), short arm (S), and total chromosome length (TL = L + S), form chromosome percentage (F% = S/ΣTL × 100), arm ratio (AR = L/S), r-value (S/L), chromosome relative length (RL% = TL/ΣTL × 100), the

TABLE 1 | Local features of indigenous harvested Iranian *Pimpinella* species were analyzed.

Species code	Local collection sites	Latitude	Longitude	Altitude (m)	Mean Temp (°C)	Mean rainfall (mm)
P. affinis (S1)	Rostamabad, Gilan	36 ° 54′ N	49 ° 21′ E	1798	19.4	1337
P. eriocarpa (S2)	Chalous, Mazandaran	36 ° 39′ N	51 ° 25′ E	29	4	785
P. tragium (S3)	Gorgan, Golestan	37 ° 28′ N	55 ° 13′ E	155	17	584
P. saxifrage (S4)	Khodaafarin, Tabriz	39 ° 13′ N	46 ° 96′ E	400	12.5	310
P. aurea (S5)	Chaldoran, West Azarbaijan	36 ° 33′ N	53 ° 03′ E	2053	28	500
P. tragioides (S6)	Bushehr	28 ° 95′ N	50 ° 83′ E	18	25	220
P. olivieri (S7)	Sarpol-e-Zahab, Kermanshah	34 ° 27′ N	45 ° 51′ E	549	30	68
P. khayyamii (S8E1)	Esfarayen, North Khorasan	37 ° 31′ N	57 ° 51′ E	1249	14.9	186
P. khayyamii (S8E2)	Ghahremanabad, North Khorasan	37 ° 20′ N	57 ° 60′ E	1800	14.9	186
P. kotschyana (S9)	Lavasanat, Tehran	35 ° 49′ N	51 ° 37′ E	1700	13.5	350
P. deverroides (S10)	Shiraz, Fars	29 ° 37′ N	52 ° 32′ E	1500	18	338
P. olivierioides (S11)	Khorramabad, Lorestan	33 ° 29′ N	48 ° 21′ E	1200	17.4	490
P. anthriscoides (S12)	Djirchal, Mazandaran	36 ° 32′ N	53 ° 09′ E	2670	15	790
P. anisactis (S13)	Bojnord, North Khorasan	37 ° 28′ N	57 ° 20′ E	1070	13.5	272
P. peucedanifolia (S14)	Urmia, West Azerbaijan	37 ° 32′ N	45 ° 04′ E	1332	12.5	450
P. khorasanica (S15)	Dargaz, Razavi Khorasan	37 ° 26′ N	59 ° 06′ E	479	14	350
P. rhodantha (S16)	Rezvanshahr, Rasht	37 ° 33′ N	49 ° 08′ E	15	16	1400

centromeric index (CI% = S/TL), and total chromosome volume (TCV = πr^2TL), where r for TCV parameter is the average chromosome radius. Idiograms were drawn based on the arm mean values, and chromosome types were recognized by the classification system of Levan et al. (1964). Furthermore, the following karyotypic asymmetry indices were measured: total form percentage (TF% = $\left(\sum S / \sum TL\right) \times 100$); Huziwara, 1962), dispersion index (DI% = (X_{CI} CV%)/100; Lavania and Srivastava, 1999), length of total chromatin (X = $2_{\sum TL}$), and total chromosome length coefficient variation (CV% = A_2 100; Paszko, 2006), difference range of relative length (DRL = RL%max-RL%min), symmetry index (S% = (TL_{min}/TL_{max}) × 100), Stebbins' (1971) classification and Romero-Zarco (1986) indexes, which are defined as the intrachromosomal asymmetry index ($A_1 = 1\text{-} \left[\sum_{i=1}^{n} (S_i / L_i) / n\right]$ and interchromosomal asymmetry index ($A_2 = sd/\overline{X}$), The mean and standard deviation are represented by x and s, respectively.

Flow Cytometric Assessment (FCM)

The 2C-DNA content of each *Pimpinella* species was estimated using flow cytometry (FCM). Flow cytometry experiments were performed using the propidium iodide (PI) staining method. A leaf of *Solanum lycopersicum* cv. Stupick with a 2C DNA value of 1.96 picograms (pg) (Doležel et al., 1992) was used as an internal reference standard. In brief, 1 cm^2 of leaves of each *Pimpinella* species, along with 1 cm^2 of the young leaves of the standard, were used to nuclei isolate by chopping with a sharp razor blade in 1 mL of woody plant buffer (Loureiro et al., 2007) in a Petri dish, supplemented with 50 $\mu g\ ml^{-1}$ propidium iodide (PI), polyvinylpyrrolidone (PVP 10) and 50 $\mu g\ ml^{-1}$ RNase. The nuclear suspension was passed through a 30 µm mesh nylon filter and then analyzed using a Cyflow Space flow cytometer (Partec GmbH, Münster, Germany) equipped with a 532 nm green high-grade solid-state laser. For each species, 5,000–10,000 nuclei per G1 peak were measured for DNA content estimates. Five different individuals per species were analyzed using linear amplification. Histograms with a coefficient of variation (CV) lower than 3% were evaluated using the FlowJo software (Version 10.6.2, Treestar, Ashland, OR, United States). Nuclear DNA content was calculated according to the following formula: Sample 2C DNA content (pg) = (Sample G_1 peak mean/standard G_1 peak mean) × standard 2C DNA amount (pg). As Doležel et al. (2003) stated, picogram values were converted to megabase pairs (Mbp), in which 1 pg of DNA represents 978 Mbp.

Statistical Analysis

The data were subjected to variance analysis (ANOVA) using the GLM procedure of the SAS software (SAS, 2003) based on a completely randomized design (CRD) with five replications for both flow cytometry and karyological data. In both cases, the normal distribution of residuals and the homogeneity of variances were approved. For mean comparisons, Tukey's test

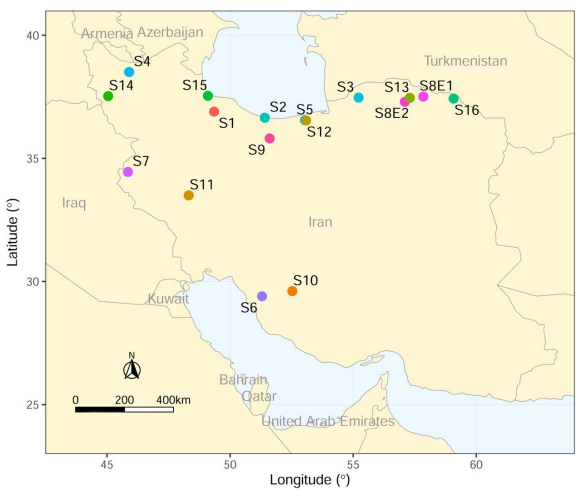

FIGURE 1 | Collection sites of Iranian endemic *Pimpinella* species on the map of Iran using ArcGIS.

was utilized (Seijo and Fernández, 2003). Multivariate statistical analysis (Srivastava, 2002) was carried out in the Minitab software package (Minitab 17, 2010) on standardized data (mean = 0, variance = 1). A principal component analysis (PCA) was performed based on a data matrix to estimate the participation of the karyotypic parameters in the species classification (Mirzaghaderi et al., 2010). Based on karyotypic parameters, cluster analysis was carried out using the unweighted pair-group method arithmetic mean (UPGMA) and the Euclidean distance (Abedi et al., 2015). The cophenetic correlation coefficient (r) was computed to specify the goodness-of-fit of the clusters to the original data. Dot plots of mean 2C-values and means of karyotypic TCV and *X* values were generated, reflecting the presence of 4C DNA in a metaphase cell during mitotic division.

RESULTS

We determined the karyotypic asymmetry indices and nuclear DNA *C*-values of 17 populations from 16 different *Pimpinella* species collected from different regions of Iran. Among 17 populations examined, 16 were diploid with varying chromosome numbers, while one [*P. rhodantha* (S16)] was tetraploid. There were significant differences

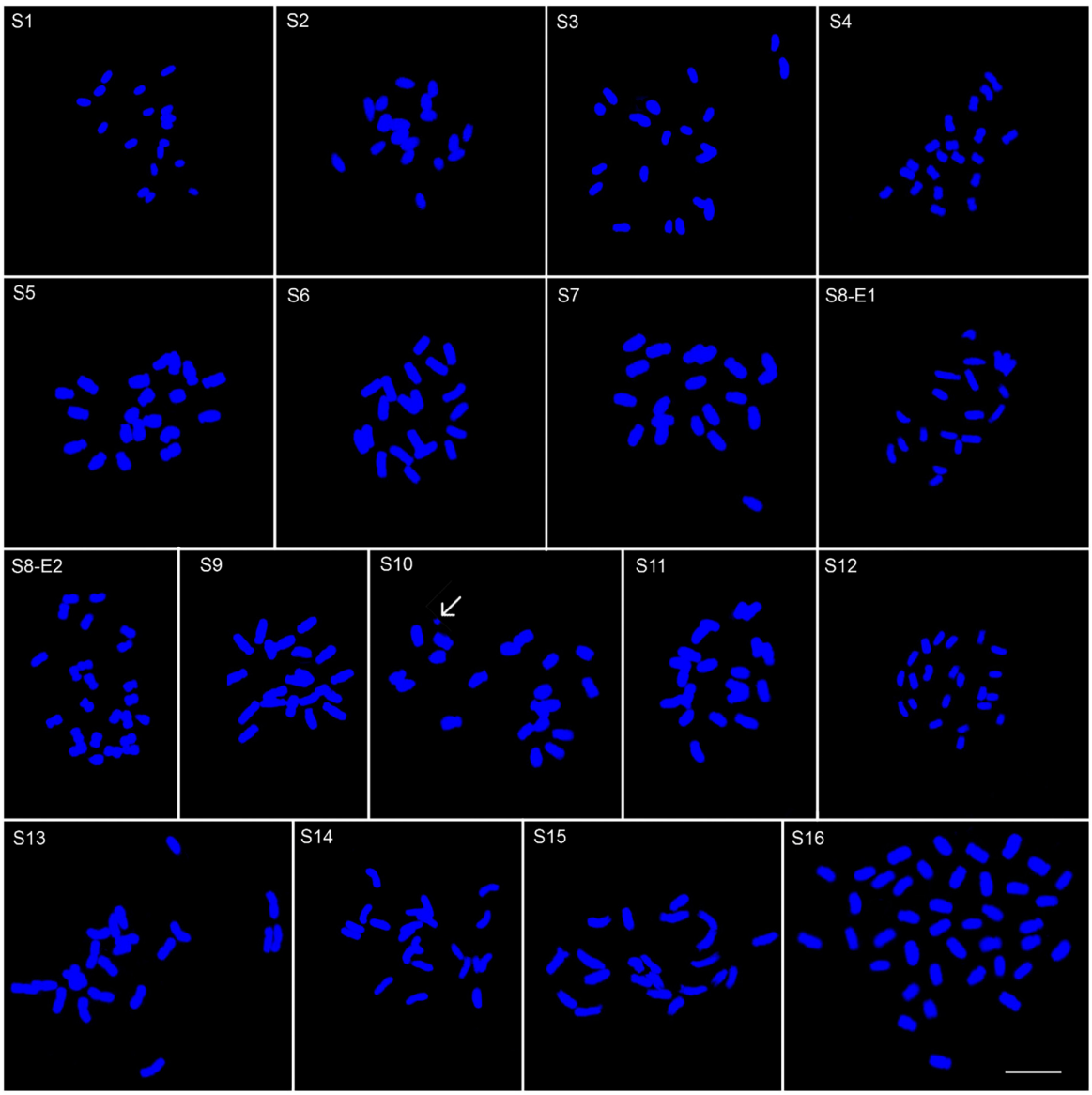

FIGURE 2 | Somatic chromosomes of 17 *Pimpinella* populations of 16 species: S1 (*P. affinis*, *x* = 9), S2 (*P. eriocarpa*, *x* = 9), S3 (*P. tragium*, *x* = 10), S4 (*P. saxifrage*, *x* = 10), S5 (*P. aurea*, *x* = 10), S6 (*P. tragioides*, *x* = 10), S7 (*P. olivieri*, *x* = 10), S8 (*P. khayyamii*, E1: *x* = 10; E2: *x* = 12), S9 (*P. kotschyana*, *x* = 10), S10 (*P. deverroides*, *x* = 10), S11 (*P. olivierioides*, *x* = 10), S12 (*P. anthriscoides*, *x* = 11), S13 (*P. anisactis*, *x* = 11), S14 (*P. peucedanifolia*, *x* = 11), S15 (*P. khorasanica*, *x* = 11), and S16 (*P. rhodantha*, *x* = 20) Scale bar indicates 10 μm. The arrow shows the B-chromosome.

among the species in their long and short arms of the chromosomes, total chromosome length, *r*-value, form percentage, total chromosome volume, and centromeric index. The figure for the mitotic metaphase complements and corresponding ideograms of the studied species are shown in **Figures 2**, **3**, respectively. In all species, the types of chromosomes were determined to be "m" (centromere at median region) and "sm," following the chromosome

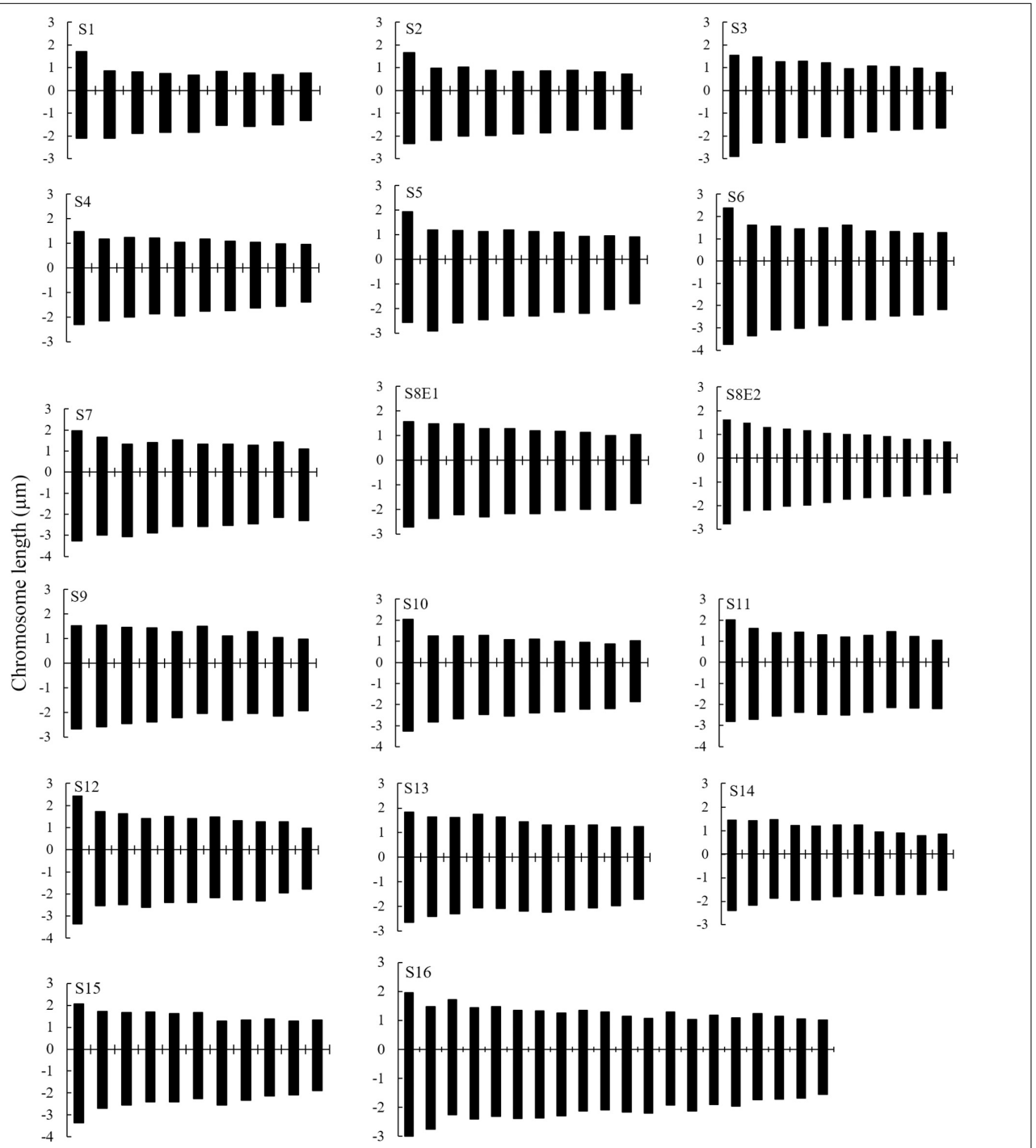

FIGURE 3 | Haploid chromosome ideograms of 17 *Pimpinella* populations of 16 species. S1 (*P. affinis*, 2m + 7sm), S2 (*P. eriocarpa*, 1m + 8sm), S3 (*P. tragium*, 5m + 5sm), S4 (*P. saxifrage*, 8m + 2sm), S5 (*P. aurea*, 1m + 9sm), S6 (*P. tragioides*, 2m + 8sm), S7 (*P. olivieri*, 3m + 7sm), S8 (*P. khayyamii*, E1: 4m + 6sm; E2: 4m + 8sm), S9 (*P. kotschyana*, 5m + 5sm), S10 (*P. deverroides*, 1m + 9sm), S11 (*P. olivierioides*, 4m + 6sm), S12 (*P. anthriscoides*, 8m + 3sm), S13 (*P. anisactis*, 11m), S14 (*P. peucedanifolia*, 7m + 4sm), S15 (*P. khorasanica*, 9m + 2sm), and S16 (*P. rhodantha*, 12m + 8sm).

TABLE 2 | The karyotypic symmetry formula for *Pimpinella* species (ST: Stebbins' type; KF: Karyotype formula).

Species	Romero-Zarco		ST	KF	DI	DRL%	CV%	S%	X	TF%
	A1	A2								
S1	0.49	0.20	3A	2m† + 7sm*	6.72	7.24	19.85	55.92	47.23	33.63
S2	0.49	0.16	3A	1m + 8sm	5.30	7.61	16.30	61.35	52.48	33.56
S3	0.42	0.19	2A	5m + 5sm	6.96	6.33	19.30	55.26	64.62	36.18
S4	0.36	0.14	1A	8m + 2sm	5.48	5.09	14.30	62.56	59.77	38.55
S5	0.48	0.16	3A	1m + 9sm	5.33	5.11	15.77	60.54	70.25	33.63
S6	0.44	0.19	2A	2m + 8sm	6.72	6.32	19.06	56.71	87.70	35.10
S7	0.45	0.14	2A	3m + 7sm	4.65	5.35	13.67	64.90	82.37	34.77
S8E1	0.40	0.13	2A	6m + 4sm	4.83	3.71	12.93	65.90	68.96	36.99
S8E2	0.50	0.18	1B	6m + 6sm	6.11	7.55	17.81	58.12	94.15	35.10
S9	0.41	0.12	2A	5m + 5sm	4.46	3.55	12.13	69.85	72.20	36.72
S10	0.52	0.19	3A	1m + 9sm	6.25	8.19	19.41	54.44	73.42	32.37
S11	0.41	0.12	2A	4m + 6sm	4.58	4.11	12.45	68.02	76.88	36.67
S12	0.36	0.20	1B	8m + 3sm	7.55	7.14	19.86	47.76	85.37	38.54
S13	0.31	0.12	1A	11m	5.02	4.44	12.45	65.80	80.35	40.66
S14	0.37	0.15	2A	7m + 4sm	5.93	6.99	15.35	63.45	66.60	37.95
S15	0.34	0.15	1A	9m + 2sm	6.16	5.03	15.48	59.60	87.54	39.21
S16	0.38	0.17	2A	12m + 8sm	6.40	3.56	16.88	50.73	137.70	37.77

*ST, Stebbins' type; KF, the karyotype formula; DI, dispersion index; DRL%, difference range of relative length; CV%, coefficient of chromosome length variation; S%, symmetry index; X, total chromatin length; and TF%, karyotype total form percentage. †m: centromere at the median region and *sm: centromere at the submedian region.*

nomenclature of Levan et al. (1964) (**Table 2**). The UPGMA grouping analysis (**Figure 4**) arrangement from this test is obtained with the PCA resulting species (**Figure 5**), which shows that the species within one cluster have the most homology in chromosomal variations. Flow cytometric data revealed a difference of 0.68 pg in 2C-value diploid populations (**Figure 6** and **Table 3**). 2C-values were correlated with and linearly regressed upon somatic metaphase, considering either TCV or X (**Figure 7**).

Chromosome Numbers and Karyotype Features

Of the 16 *Pimpinella* species examined, two [*P. affinis* (S1) and *P. eriocarpa* (S2)] had chromosome numbers of $2n = 2x = 18$ with a basic chromosome number $x = 9$, eight [*P. tragium* (S3), *P. saxifrage* (S4), *P. aurea* (S5), *P. tragioides* (S6), *P. olivieri* (S7), *P. kotschyana* (S9), *P. deverroides* (S10) and *P. olivierioides* (S11)] were determined as $2n = 20$ with a basic chromosome number $x = 10$, four [*P. anthriscoides* (S12), *P. anisactis* (S13), *P. peucedanifolia* (S14), and *P. khorasanica* (S15)] were determined as $2n = 22$, and one [*P. rhodantha* (S16)] was tetraploid ($2n = 4x = 40$). Interestingly, two chromosome numbers, $2n = 2x = 20$ and 24, were observed in the *P. khayyamii* (S8) population.

The chromosome numbers were determined as $2n = 18$, 20, 22, 24, and 40, with a basic chromosome number of $x = 9$, 10, 11, 12, and two diploid and tetraploid levels. The frequencies of the observed chromosome numbers were highly varied. Most of the species (52.94%) showed $2n = 2x = 20$; 11.76% were $2n = 2x = 18$; 23.53% were $2n = 2x = 22$; 5.88% were $2n = 2x = 24$; and 5.88% of the species were $2n = 4x = 40$. The sporophytic chromosome numbers ($2n$) and karyotypic details for the studied species are presented in **Table 2**.

The size of the short and long arms, chromosomal length, *r*-value, TF %, total chromosome volume (TCV), arm ratio (AR), and centromeric index (CI) differed significantly across the species. The average chromosomal length varies from 2.62 μm (S1) to 4.38 μm (S6), and the haploid genome length varies from 23.62 μm (S1) to 43.85 μm (S6). The CI mean of the diploid supplement ranged between 33.10% (S2) and 40.0 (S13). In the tetraploid species (*P. rhodantha*), the mean TL, the haploid genome length, and CI were 3.44 μm, 68.85 μm, and 37.68%, respectively. Considering mean values, species with $2n = 18$ have the smallest TL and CI. The species with $2n = 20$ have an intermediate CI but the highest TL (3.64 μm), whereas those with $2n = 22$ have the highest CI but a TL similar to that of $2n = 20$. **Figure 8** depicts the TL and CI correlations for each species. Using Levan et al. (1964) chromosome classification, two chromosome types of "m" (The centromere is located in the central region.) and "sm" (The centromere is located in the subcentral region.) formed 14 different karyotypic formulas. **Table 2** lists the various types and numbers of karyotypic formulas/species. According to different karyotypic symmetrical indexes tested, the *Pimpinella* species studied showed various symmetrical karyotypes. The greatest TF percentage value was found in S13 (40.66 %, the most symmetric), while the lowest was found in S10 (32.37%, the most asymmetric). However, the highest value of S% was identified in S9 (69.85%, the most symmetric), while S12 had the lowest value (47.76%, the most asymmetric). The highest and the lowest values of DRL% were distinguished in S10 (8.19%, the most asymmetric) and S9 (3.55%, the most symmetric), respectively. The highest and the lowest values of CV% belonged to S12 (19.86%, the most

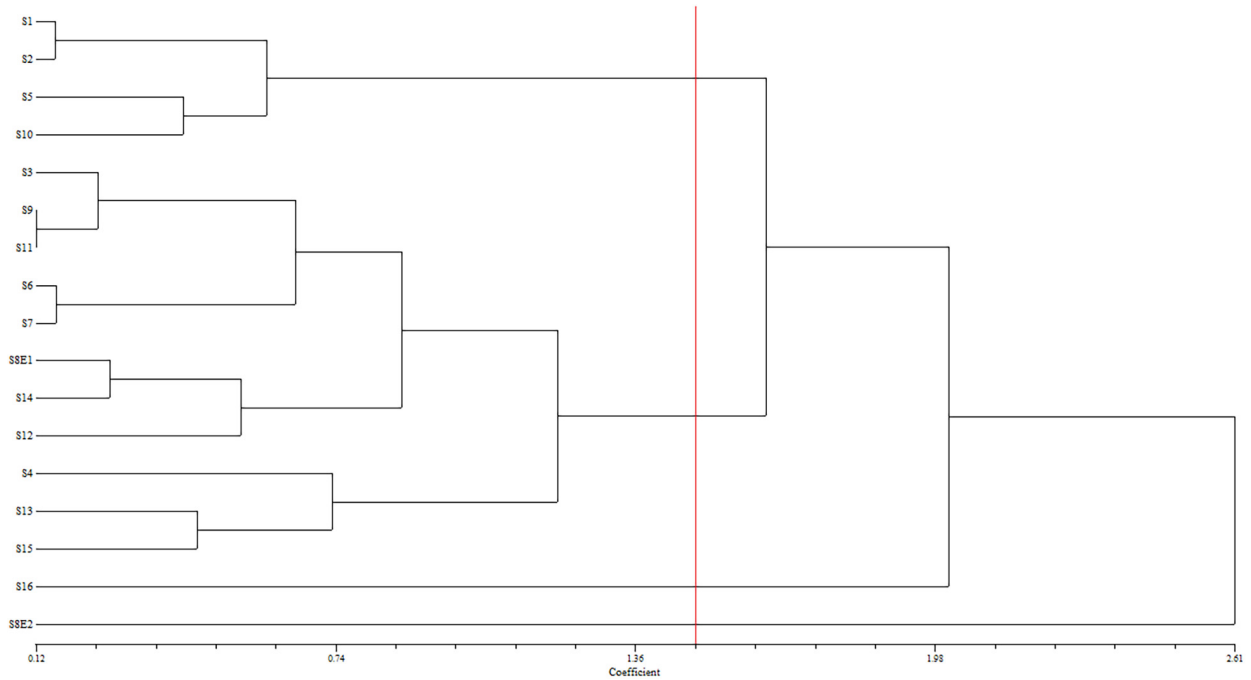

FIGURE 4 | Dendrogram showing the phenetic relationships among the examined populations of *Pimpinella* of 16 species, created using the matrix of karyotype similarities with UPGMA; cophenetic correlation $r = 0.80$. S1 (*P. affinis*, $x = 9$), S2 (*P. eriocarpa*, $x = 9$), S3 (*P. tragium*, $x = 10$), S4 (*P. saxifrage*, $x = 10$), S5 (*P. aurea*, $x = 10$), S6 (*P. tragioides*, $x = 10$), S7 (*P. olivieri*, $x = 10$), S8 (*P. khayyamii*, E1: $x = 10$; E2: $x = 12$), S9 (*P. kotschyana*, $x = 10$), S10 (*P. deverroides*, $x = 10$), S11 (*P. olivierioides*, $x = 10$), S12 (*P. anthriscoides*, $x = 11$), S13 (*P. anisactis*, $x = 11$), S14 (*P. peucedanifolia*, $x = 11$), S15 (*P. khorasanica*, $x = 11$), and S16 (*P. rhodantha*, $x = 20$).

asymmetric) and S9 (12.13%, the most symmetric), respectively. Similar to the CV percent results, the greatest DI value was observed in S12 (7.55 %, the most asymmetric), while the lowest value was found in S13 (5.02%, the most symmetric) (**Table 2**). In conclusion, four (S%, DRL%, CV%, and DI%) among the five karyotypic symmetrical groups tested confirmed that among all 17 *Pimpinella* populations examined, S12 and S9 appear to have the most asymmetrical and symmetrical karyotypes, respectively.

Three species (S4, S13, and S15) were classified as 1A; eight species (S3, S6, S7, S8E1, S9, S11, S14, and S16) were classified as 2A; four species (S1, S2, S5, and S10) were classified as 3A; and the karyotype of two species (S12 and S8E2) was classified as 1B, according to the Stebbins classification (Stebbins, 1971; **Table 2**). The four groups of species are shown in the scatter plot of the A1 and A2 asymmetry indices: (1) taking into account one species (S13) with the most symmetrical karyotypes (A1 and A2 average = 215), (2) comprising seven species (S4, S8E1, S9, S11, S14, S15, and S16) with the symmetrical karyotypes (A1 and A2 average = 0.260), (3) including three species (S3, S6, and S7) with the asymmetrical karyotypes (A1 and A2 average = 0304), and (4) including five species (S1, S2, S5, S8E2, and S10) with the most asymmetrical karyotypes (A1 and A2 average = 0337, **Figure 9**). S12 falls between the second and third groups. Both indices differentiate S13 from other species, and the A1 index discriminates species by base chromosome number.

Other interesting peculiarities were found in the meta-phase plate of *P. deverroides* (S10). The chromatin body of **Figure 2** could account for one acentric fragment, with the chromatids lying parallel through their length. However, another acentric fragment – corresponding to the broken chromosome end of the other fused chromosome – should also be visible, but in this case,

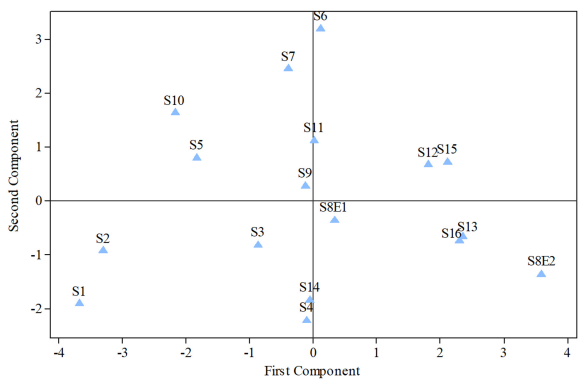

FIGURE 5 | Diagram produced from the examination of the first and second components of *Pimpinella* species (strongly linked to the placement of the centromere and the length of the complements, accordingly). S1 (*P. affinis*), S2 (*P. eriocarpa*), S3 (*P. tragium*), S4 (*P. saxifrage*), S5 (*P. aurea*), S6 (*P. tragioides*), S7 (*P. olivieri*), S8 (*P. khayyamii*), S9 (*P. kotschyana*), S10 (*P. deverroides*), S11 (*P. olivierioides*), S12 (*P. anthriscoides*), S13 (*P. anisactis*), S14 (*P. peucedanifolia*), S15 (*P. khorasanica*), and S16 (*P. rhodantha*).

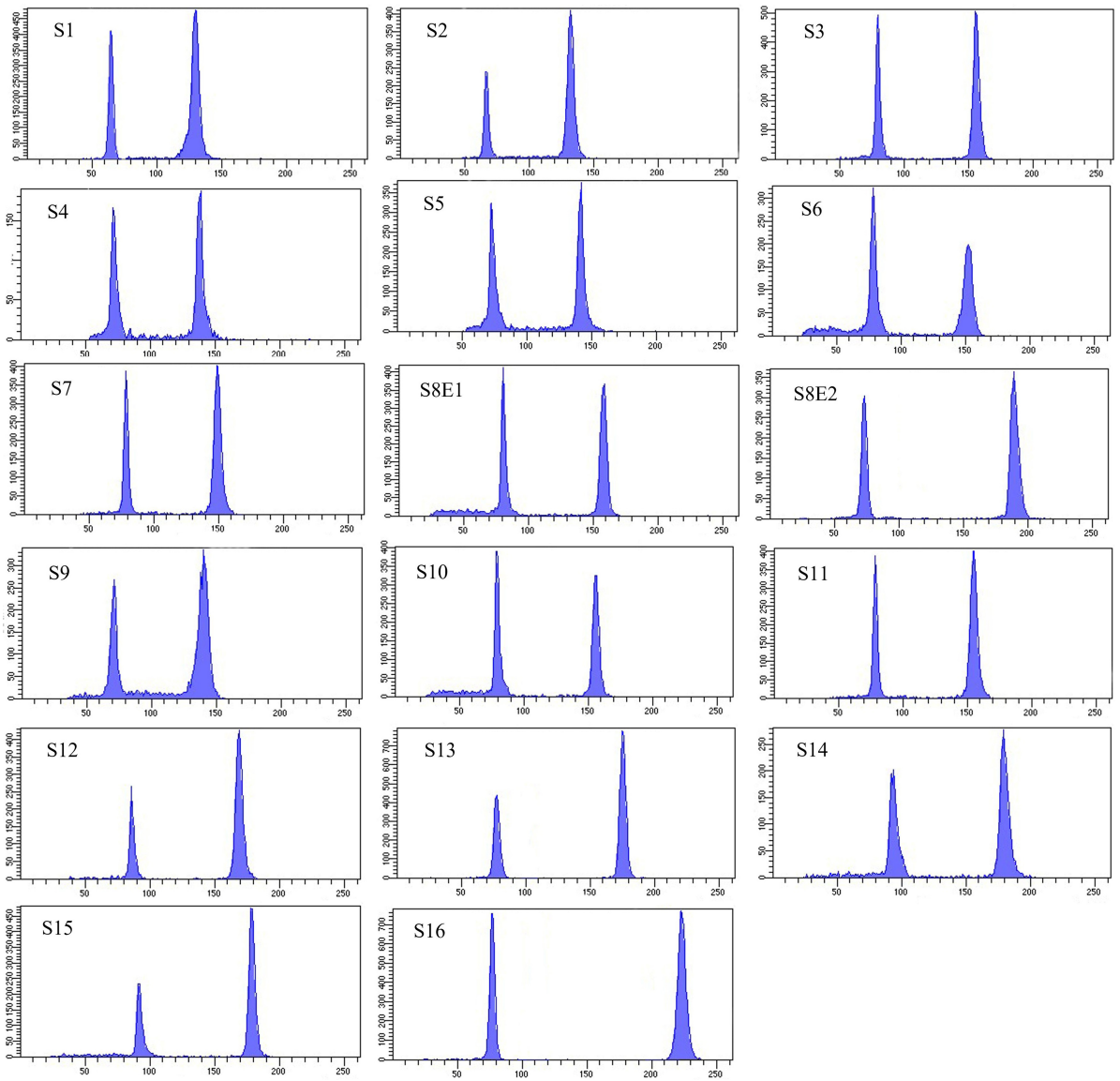

FIGURE 6 | Histograms of flow cytometric 2C DNA content of 17 populations of 16 *Pimpinella* species. The left peaks refer to the *Solanum lycopersicum* cv. Stupick (2C DNA = 1.96 pg) internal reference standard and the right peaks to the *Pimpinella* samples. The *x*-axes are fluorescence intensity and the number of nuclei, respectively. S1 (*P. affinis*), S2 (*P. eriocarpa*), S3 (*P. tragium*), S4 (*P. saxifrage*), S5 (*P. aurea*), S6 (*P. tragioides*), S7 (*P. olivieri*), S8 (*P. khayyamii*), S9 (*P. kotschyana*), S10 (*P. deverroides*), S11 (*P. olivierioides*), S12 (*P. anthriscoides*), S13 (*P. anisactis*), S14 (*P. peucedanifolia*), S15 (*P. khorasanica*), and S16 (*P. rhodantha*).

it is not. Hence, another explanation for this chromatin body would be that it is a B-chromosome.

The UPGMA dendrogram – constructed using the matrix of karyotype similarities (**Figure 4**) – displayed four major clusters. The first cluster is comprised of S1, S2, S5, and S10, distinguished by the shortest complements but high RL and AR. In this cluster, species with 18 chromosomes form a subgroup. The second cluster comprises 11 species with both 20 and 22 chromosomes, which are characterized by a high L, F%, and TL. The S16 species, with 40 chromosomes with the lowest F%, form the third cluster. The fourth cluster contained S8E2, characterized by a high S, CI%, *r*-value, and the lowest AR.

The first two PCs in the karyotypic parameter's principal component analysis (PCA) account for 80.8% of the cumulative variation, and they were shown in a 2-dimensional image (**Figure 5**). The first component (50.9%) emphasizes the position of the centromere, while the second component (29.9%) accentuates variation in complement length. The resulting species arrangement from this test entirely matches with those obtained from the UPGMA clustering method.

Seventeen populations of 16 *Pimpinella* species were analyzed using FlowJo software version 10.6.2 to calculate the amount of DNA in the nucleus. The acquired histograms for estimating the nuclear DNA content of each species contained two peaks;

the right peaks refer to the known internal reference standard (*Solanum lycopersicum* cv. Stupick), and the left peaks refer to the unknown *Pimpinella* species (**Figure 6**). **Table 3** shows DNA content (pg) and the genome sizes (Mbp) of studied *Pimpinella* species. The 2C-value varied from 2.48 to 5.50 pg (relating to a diploid *P. affinis* and a tetraploid *P. rhodantha*, respectively). Among the 16 diploids with chromosome numbers ranging from 18 to 24, a distinction of 0.68 pg in 2C-value (2.48 and 3.16) was observed (**Table 3**), while the mean 2C-value of a *P. rhodantha* ($2n$ = 40, 5.50 pg) was determined to be precisely double the value of the two diploids with x = 10 ($2n$ = 20, 2.75 pg). Tukey's test revealed a significant difference in the 2C-value among the 16 diploid species. Meanwhile, the holoploid genome size ranged from 1212.72 Mbp (diploid S1) to 1511.01 Mbp (diploid S13), with a difference of 298.29 Mb, whereas the tetraploid haploid genome size was 2870.43 Mbp. Significant correlation was observed between chromosomal features measured, viz, X, and TCV (r = 0.879** and 0.701**, respectively, in **Figures 7A,B**), with 2C-values demonstrating linear relationships (b = 0.034** and 1.897**, in **Figures 7A,B**, respectively).

DISCUSSION

The findings of this study provide precise photos of the chromosomal characteristics of indigenous Iranian *Pimpinella* species for the first time. There is little information available on the chromosome numbers. The karyotype data of the studied species and information on the 2C DNA amount data were utterly insufficient. The findings of the present study revealed two ploidy levels: diploids ($2x$) and tetraploids ($4x$), as well as four different chromosomal numbers: 18, 20, 22, and 40. In prior research, such chromosome counts of 18 (Yurtseva, 1988; Pimenov et al., 2003), 20 (Al-Eisawi, 1989; Verlaque and Filosa, 1992; Pimenov et al., 1996), 22 (Constance and Chuang, 1982; Abebe, 1992; Castro and Rossello, 2007), 40 (Daushkevich et al., 1995), and ploidy levels have also been reported for other species of the *Pimpinella* genus. Furthermore, *P. affinis* was found to have different chromosome numbers of 16 and 18 (Jurtseva, 1988), *P. tragium* had 20 and 22 (Galland, 1988), *P. eriocarpa* showed 16 (Al-Eisawi, 1989), *P. rhodanth* displayed 18 and 20 (Daushkevich et al., 1995), and *P. saxifrage* showed $2n$ = 18, 20, 36, and 40 (Gawlowska, 1967).

The differences in chromosome number and chromosome morphology found among species indicate that chromosome structural changes may be used to distinguish species that are very similar to each other and cannot be separated using morphological characters. Among these species, *P. khorasanica* differs from *P. anisactis* due to the presence of two "sm" chromosome types and possesses a shorter total chromatin length. These relative differences could be used for breeding programs by facilitating chromosome identification in hybrid populations and derivatives in *Pimpinella*. Parent combinations with relative differences in the chromosome numbers/type due to chromosome pairing could result in a successful cross (Hamidi et al., 2018; Akbarzadeh et al., 2021). Three species – S3 (*P. tragium*), S13 (*P. anisactis*), and S15 (*P. khorasanica*) – are very similar to each other from the viewpoint of anatomy, and we are not able to separate them with anatomical features. In addition, they are morphologically similar to each other. Therefore, the findings of this study can be considered as a guide to separating these species, especially S13 (*P. anisactis*) and S15 (*P. khorasanica*), which are distributed in a small area in Khorassan, where they are endemic. Taxonomic criteria, including chromosome number and asymmetric indices, have been used in plant phylogenetic and taxonomic consideration of the genus (Hesamzadeh Hejazi and Ziaei Nasab, 2010).

TABLE 3 | 2C-value and flow cytometric DNA estimation data of *Pimpinella* species.

Species	Ploidy level	2n	2C-value (pg ± SE)	1C-value (pg)	1Cx-value (pg)	Holoploid genome size (Mbp)	Monoploid genome size (Mbp)
S1	2x	18	2.48^{j} ± 0.021	1.240	1.240	1212.72	1212.72
S2	2x	18	2.55^{hj} ± 0.018	1.275	1.275	1246.95	1246.95
S3	2x	20	2.85^{de} ± 0.018	1.425	1.425	1393.65	1393.65
S4	2x	20	2.62^{ghj} ± 0.021	1.310	1.310	1281.18	1281.18
S5	2x	20	2.66^{fgh} ± 0.020	1.330	1.330	1300.74	1300.74
S6	2x	20	2.79^{ef} ± 0.022	1.395	1.395	1364.31	1364.31
S7	2x	20	2.71^{efg} ± 0.022	1.355	1.355	1325.19	1325.19
S8E1	2x	20	2.83^{de} ± 0.022	1.415	1.415	1383.87	1383.87
S8E2	2x	24	3.16^{b} ± 0.017	1.580	1.580	1545.24	1545.24
S9	2x	20	2.70^{efgh} ± 0.020	1.350	1.350	1320.30	1320.30
S10	2x	20	2.84^{de} ± 0.021	1.420	1.420	1388.76	1388.76
S11	2x	20	2.80^{ef} ± 0.032	1.400	1.400	1369.20	1369.20
S12	2x	22	2.98^{cd} ± 0.022	1.490	1.490	1457.22	1457.22
S13	2x	22	3.09^{bc} ± 0.025	1.545	1.545	1511.01	1511.01
S14	2x	22	3.05^{bc} ± 0.021	1.525	1.525	1491.45	1491.45
S15	2x	22	3.08^{bc} ± 0.021	1.540	1.540	1506.12	1506.12
S16	4x	40	5.50^{a} ± 0.021	2.750	1.375	2689.50	1344.75

According to Tukey's test, means with different symbol letters in columns are significantly different ($P < 0.01$). Means with the same letter are not statistically different ($P > 0.05$).

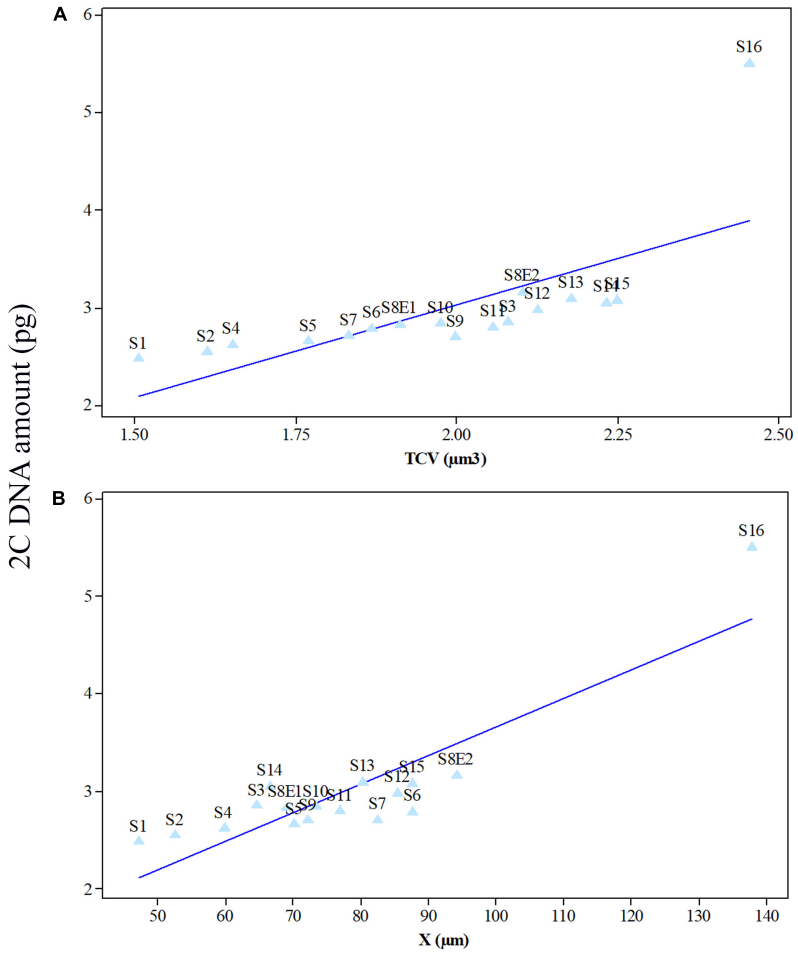

FIGURE 7 | Mean leaf 2C DNA amounts (pg) of 17 *Pimpinella* populations of 16 species plotted against mean meristem total chromosome volumes (TCV, μm^3) **(A)** and chromatin lengths (X, µm) **(B)** at mitotic metaphase. S1 (*P. affinis*), S2 (*P. eriocarpa*), S3 (*P. tragium*), S4 (*P. saxifrage*), S5 (*P. aurea*), S6 (*P. tragioides*), S7 (*P. olivieri*), S8 (*P. khayyamii*), S9 (*P. kotschyana*), S10 (*P. deverroides*), S11 (*P. olivierioides*), S12 (*P. anthriscoides*), S13 (*P. anisactis*), S14 (*P. peucedanifolia*), S15 (*P. khorasanica*), and S16 (*P. rhodantha*).

Among species with 22 chromosomes, *P. anthriscoides* can be clearly distinguished from other species (*P. khorasanica*, *P. peucedanifolia*, and *P. anisactis*) by different evolutionary karyotype classification and karyotype formulas. This species displayed a lower 2C-value in comparison with other species with the same number of chromosomes. Also, *P. anthriscoides* differs from other species of the genus *Pimpinella* in some morphological traits, such as plant height, tepal and leaf area, and the size of reproductive organs (data not shown). Based on these observations, *P. anthriscoides* is introduced as a separate species in the distinct genus *Pseudopimpinella* in this report. Changes in chromosome morphology and genome size have been developed and used as the basic mechanisms in plant taxonomy and phylogenetic consideration of the genus (Bernardos et al., 2003; Navarro et al., 2004; Arslan et al., 2012).

To the best of our knowledge, there has been no cytological report of the presence of the B-chromosome in *Pimpinella* species. Hence, this study is the first report in this genus. The B-chromosomes are extra chromosomes and smaller than the usual A-chromosomes, of which the origin and functions are not well known (Palestis et al., 2004; Pellicer et al., 2007). The presence of B-chromosomes has been reported in plant taxa (Fregonezi et al., 2004; Felix et al., 2011; Abedi et al., 2015), and they are not necessarily for the survival of the species; however, they may act in either a positive or negative role as an adaptive function or parasitic genome, respectively (Pellicer et al., 2007; Jones, 2012). The recognition of the B-chromosome in a few individuals may favor the hypothesis of a parasitic B-chromosome (Felix et al., 2011). In the present study, the persistent presence of a B-chromosome in all examined individuals of S10 seems to support the hypothesis of an adaptive function of the B-chromosome.

Another interesting novel finding in this study was the first record of 24 chromosomes in an Iranian *P. khayyamii* (S8E2). So far, no other *Pimpinella* species have been identified with this chromosome number. Interestingly, the S8E2 *Pimpinella* population had four more chromosomes than the other diploid Iranian endemic *Pimpinella* population (S8E1; Esfarayen, North

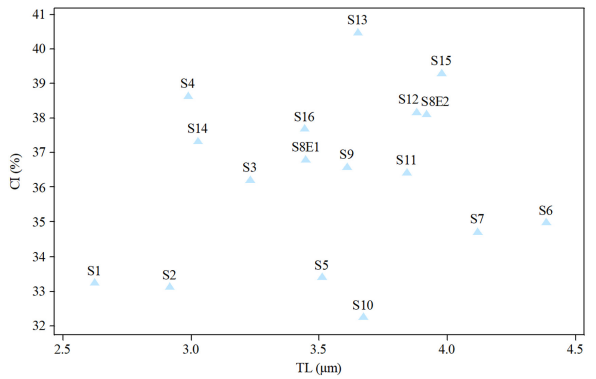

FIGURE 8 | Relationship between the total length of haploid complement (TL) and centromeric index (CI) of 17*Pimpinella* populations of 17 species.

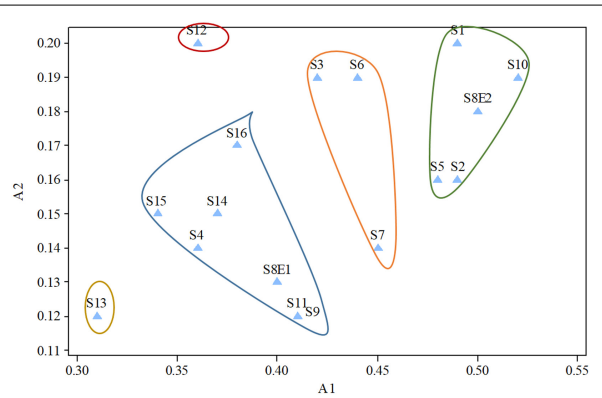

FIGURE 9 | Two-dimensional graphic of the Romero-Zarco asymmetry indices (A1: intrachromosomal asymmetry index, A2: interchromosomal asymmetry index) of 16 species (17 population) *Pimpinella* species. S1 (*P. affinis*), S2 (*P. eriocarpa*), S3 (*P. tragium*), S4 (*P. saxifrage*), S5 (*P. aurea*), S6 (*P. tragioides*), S7 (*P. olivieri*), S8 (*P. khayyamii*), S9 (*P. kotschyana*), S10 (*P. deverroides*), S11 (*P. olivierioides*), S12 (*P. anthriscoides*), S13 (*P. anisactis*), S14 (*P. peucedanifolia*), S15 (*P. khorasanica*), and S16 (*P. rhodantha*).

Khorasan, Iran, **Figure 1**), which was collected from the same geographic regions with only small differences in longitude and latitude. Previous research and our findings may lead us to conclude that the instability in both chromosome number and ploidy levels in studied *Pimpinella* species is probably due to interspecific hybridization and polyploidization, which, in turn, induces a cascade of subsequent genomic rearrangements. Above all, epigenomic rearrangements (Maletskii, 2004) may also lead to epigenetic silencing (Scheid et al., 1996). These rearrangements may increase the adaptive capacity of certain species.

According to the chromosomal parameters measured in the current study, the two diploid species with 18 chromosomes, the nine with 20 chromosomes, and the four with 22 chromosomes demonstrated intra- and inter-specific variation in X, TL, and TCV. Considering the primary chromosome parameter, S6 – which was geographically isolated from other species – appears to exhibit the largest chromatin length (X) among other *Pimpinella* species, either diploids with different chromosome numbers or tetraploids. This might indicate that chromosomal length is affected by geographical and environmental adaptability. In spite of the observed intra- and inter-specific variation, the bulk of karyotypic symmetrical indices suggests that most *Pimpinella* species, including diploids and tetraploids, possess symmetric and primitive karyotypes, which is most likely due to inter/intra hybridization and polyploidization. Thus, their similar karyotype structure causes their tendency toward crosses and does not cause a disturbance in reproduction. In general, it is believed that asymmetric karyotype can be linked to the evolutionary history of a particular group of plants (Stebbins, 1971). The high value of the A1 index (variable between 0 and 1) is considered a specialized adaptation, whereas the interchromosomal asymmetry index (A2, variable between 0 and ∞) is associated with the relative taxonomic distance between species of different taxa (Romero-Zarco, 1986). The species *P. anisactis* (S13) had the smallest A1 and A2 index values, which are probably attributable to its strong adaptability to its habitat conditions; it is therefore not particularly specialized. Three species, S1, S10, and S8E2, had the largest values of the A1 and A2 indices, indicating that these may be well-specialized species.

According to our results, FCM was effectively conducted to analyze ploidy level stability of species (Wyman et al., 1992). Different Iranian *Pimpinella* species were separated based on their nuclear DNA content, indicating inter/intraspecies diversity and confirming the cytological findings. Variability in DNA C−values is a prerequisite for use as a taxonomic character (Ellul et al., 2002). Previously, there was only a single report of the 2C-value in the diploid *P. saxifrage* (2C DNA = 8.52 pg, Temsch et al., 2010). However, this value differs considerably from our data. The reason for this is unknown but could arise from the cell cycle, rate of cell division, radiation sensitivity, ecological demeanor in plant societies and life forms, and differences between methods of DNA content analysis (Bennett et al., 2000). A surprising finding related to the 2C value is that the two diploid *P. khayyamii* accessions (S8E1, S8E2) with two different chromosome numbers of $2n = 2x = 20$, 24 and two different 2C DNA content were gathered from the same area (North Khorasan) with a slight difference in geographical coordinates (**Table 1**). A major question remains unanswered: what causes *P. khayyamii* from the same geographic regions to differ in four chromosomes? Are there any genetic or ecological attributes that lead to this difference? Systematic investigation into different aspects of this needs to be undertaken.

The mean comparison of 2C value/chromosome between 16-diploids (0.138 pg) and only tetraploid (0.137 pg) was not substantially different (t-value = 0.00032; P-value = 0.74). For monoploid genome size, a similar conclusion was obviously true (13,811 Mbp for diploids, 1344.75 Mbp for tetraploid; t-value = 1.470; P-value = 0.163). In other words, our finding indicated that the 2C DNA-value mean and 1Cx genome size mean in the examined *Pimpinella* species were not proportional to ploidy level and the nuclear DNA content per basic chromosome set (1 Cx) tended to decrease when a high ploidy level was observed. A broad analysis of the mean basic genome size at different ploidies in angiosperms showed that, while basic genome size decreased with increasing ploidy, those with

larger mean genome sizes at the diploid level showed a greater reduction than those with smaller mean genome sizes (Kellogg and Bennetzen, 2004; Leitch and Bennett, 2004). Studying 67 *Artemisia* species with different ploidy levels showed that 1C*x* genome size tended to decrease significantly in polyploids compared with diploids (Pellicer et al., 2007). Exceptions to this are found in the bromegrass germplasm accessions, where only a slight reduction of DNA content was detected as the ploidy level increased (Tuna et al., 2001). Different patterns of genome size variation linked to different kinds of evolutionary mechanisms and the nature of speciation/polyploidization have been shown in previous studies (Bennetzen and Kellogg, 1997; Soltis et al., 2003). Genome size reduction mechanisms, along with allopolyploidization in *Aegilops,* could be an obligatory adaptation in polyploid genome evolution (Ozkan et al., 2003). Hence, polyploidy is one possible contributor to *C*-value variation, but the relationship between *C*-value and ploidy is not straightforward (Leitch and Bennett, 2004; Murray et al., 2005). In this study, there is no relationship between the 2C DNA content of tetraploids and diploids in *Pimpinella*. Our karyotypic data on *Pimpinella* species showed a karyotype formula of 6m + 4sm for diploid S8E1 and 6m + 6sm for S8E2. In the karyotype of S8E1, metacentric "m" chromosomes were predominant, and S8E2 varied in the two types of submetacentric "sm" chromosomes. This may help us to deduce that the newly reported 24 chromosomes diploid *Pimpinella* tends to have a different evolutionary karyotype classification from the "2A" Stebbins karyotype category (relative symmetric karyotype) for most diploids to the "1B" relatively asymmetric karyotype.

In *Pimpinella* species, the significant positive correlation between the 2C-value and some karyotypic features, including the ploidy level, the total length of chromatin, and total chromosome volume, indicates that changes in nuclear DNA content have accompanied chromosome structural changes. In agreement with our finding, such a relationship between 2C-value and chromosomal parameters has been reported in *Tulipa* (Abedi et al., 2015); *Lathyrus* (Karimzadeh et al., 2011), *Thymus* (Mahdavi and Karimzadeh, 2010), and *Helichrysum* (Azizi et al., 2014).

Interestingly, in nine species with 20 chromosomes, the 1C*x* genome size of S4 (1C*x* = 1281.18 Mbp) was considerably lower (8.07%, $P < 0.05$) than the other diploid S3 (1393.65 Mbp). The average 1C*x* genome size of S3 was more than that of either diploids or tetraploids (S16, 1C*x* = 1344.75 Mbp), giving us an expanded view of variation among these species. Such considerable variability could be attractive for either polyploid induction or hybrid production among types of studied *Pimpinella* species. As a result, genome size might be utilized as an effective marker for detecting hybrids (Ellul et al., 2002).

Further work, such as fluorescence *in situ* hybridization (FISH), using repetitive sequences and rDNA genes and C-banding complementary assessment of karyology and cytology in meiosis would add more data to *Pimpinella* taxonomic studies. Our data, in combination with additional data on *Pimpinella* species worldwide, would help to recognize the origin and evolution of this genus and help to protect the endemic and threatened species with different ploidy levels and chromosome numbers. Moreover, knowledge about genome size is helpful in demonstrating any relationship between nuclear DNA content and the ecological niches of *Pimpinella* species.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

SM, GM, JB, and GR conceived and designed this study. SM and AS-E conducted the experiments. SM and MH analyzed the data. SM wrote the manuscript. HN-Z and DE revised the manuscript. All authors have read and approved the published version of the manuscript.

FUNDING

This research was supported by the University of Western Australia, WA, Australia.

REFERENCES

Abdolmalaki, Z., Mirzaghaderi, G., Mason, A. S., and Badaeva, E. D. (2019). Molecular cytogenetic analysis reveals evolutionary relationships between polyploid *Aegilops* species. *Plant Syst. Evol.* 305, 459–475. doi: 10.1007/s00606-019-01585-3

Abebe, D. (1992). Systematic studies in the genus *Pimpinella* L. (Umbelliferae) from tropical Africa. *Bot. J. Linn. Soc.* 110, 327–372. doi: 10.1111/j.1095-8339.1992.tb00298.x

Abedi, R., Babaei, A., and Karimzadeh, G. (2015). Karyological and flow cytometric studies of *Tulipa* (Liliaceae) species from Iran. *Plant Syst. Evol.* 301, 1473–1484. doi: 10.1007/s00606-014-1164-z

Akbarzadeh, M., Van Laere, K., Leus, L., De Riek, J., Van Huylenbroeck, J., Werbrouck, S. P., et al. (2021). Can knowledge of genetic distances, genome sizes and chromosome numbers support breeding programs in hardy geraniums? *Genes* 1:730. doi: 10.3390/genes12050730

Al-Eisawi, D. M. (1989). Chromosome counts of Umbelliferae of Jordan. *Ann. Bot.* 47, 201–214.

Arslan, E., Ertuğrul, K., and Öztürk, A. B. (2012). Karyological studies of some species of the genus *Vicia* L. (Leguminosae) in Turkey. *Caryologia* 65, 106–113. doi: 10.1080/00087114.2012.709804

Azizi, N., Sheidai, M., Mozaffarian, V., and Nourmohammadi, Z. (2014). Karyotype and genome size analyses in species of *Helichrysum* (Asteraceae). *Acta Bot. Bras.* 28, 367–375. doi: 10.1590/0102-33062014abb3136

Bainard, J. D., Forrest, L. L., Goffinet, B., and Newmaster, S. G. (2013). Nuclear DNA content variation and evolution in liverworts. *Mol. Phylogenet. Evol.* 68, 619–627. doi: 10.1016/j.ympev.2013.04.008

Bancheva, S., and Greilhuber, J. (2006). Genome size in Bulgarian *Centaurea* s.l. (Asteraceae). *Plant Syst. Evol.* 257, 95–117. doi: 10.1007/s00606-005-0384-7

Bennett, M. D., Bhandol, P., and Leitch, I. J. (2000). Nuclear DNA amounts in angiosperms and their modern uses- 807 new estimates. *Ann. Bot.* 86, 859–909. doi: 10.1006/anbo.2000.1253

Bennett, M. D., and Leitch, I. J. (2000). *Variation in Nuclear DNA Amount (Cvalue) in Monocots: Systematics and Evolution*. Mebourne, VIC: CSIRO, 137–146.

Bennett, M. D., Price, H. J., and Johnston, J. S. (2008). Anthocyanin inhibits propidium iodide DNA fluorescence in *Euphorbia pulcherrima*: implications for genome size variation and flow cytometry. *Ann. Bot.* 101, 777–790. doi: 10.1093/aob/mcm303

Bennetzen, J. L., and Kellogg, E. A. (1997). Do plants have a one-way ticket to genomic obesity? *Plant Cell* 9, 1509–1514. doi: 10.1105/tpc.9.9.1509

Bernardos, S., Amich, F., and Crespí, A. (2003). Karyological and taxonomical notes on three species of the genus *Epipactis* (*Neottioideae*, *Orchidaceae*) in the central-western Iberian Peninsula. *Folia Geobot.* 38, 319–331. doi: 10.1007/BF02803202

Castro, M., and Rossello, J. A. (2007). Karyological observations on plant taxa endemic to the Balearic Islands. *Bot. J. Linn. Soc.* 153, 463–476. doi: 10.1111/j.1095-8339.2007.00617.x

Constance, L., and Chuang, T. I. (1982). Chromosome numbers of Umbelliferae (Apiaceae) from Africa south of the Sahara. *Bot. J. Linn. Soc.* 85, 195–208. doi: 10.1111/j.1095-8339.1982.tb02586.x

Daushkevich, J. V., Alexeeva, T. V., and Pimenov, M. G. (1995). IOPB chromosome data 10. *Int. Org. Plant Biosyst. Newsl.* 25, 7–8.

Dobigny, G., Ducroz, J. F., Robinson, T. J., and Volobouev, V. (2004). Cytogenetics and cladistics. *Syst. Biol.* 53, 470–484. doi: 10.1080/10635150490445698

Doležel, J., Bartos, J., Voglmayr, H., and Greilhuber, J. (2003). Nuclear DNA content and genome size of trout and human. *Cytom. J. Int. Soc. Anal. Cytol.* 51, 127–128. doi: 10.1002/cyto.a.10013

Doležel, J., Greilhuber, J., and Suda, J. (2007). Estimation of nuclear DNA content in plants using flow cytometry. *Nat. Protoc.* 2, 2233–2244. doi: 10.1038/nprot.2007.310

Doležel, J., Sgorbati, S., and Lucretti, S. (1992). Comparison of three DNA fluorochromes for flow cytometric estimation of nuclear DNA content in plants. *Physiol. Plant.* 85, 625–631. doi: 10.1111/j.13993054.1992.tb04764.x

Downie, S. R., Spalik, K., Katz-Downie, D. S., and Reduron, J. P. (2010). Major clades within Apiaceae subfamily Apioideae as inferred by phylogenetic analysis of nrDNA ITS sequences. *Plant Div. Evol.* 128, 111–136. doi: 10.1016/j.ympev.2021.107183

Ellul, P., Boscaiu, M., Vicente, O., Moreno, V., and Rossello, J. A. (2002). Intra- and interspecific variation in DNA content in *Cistus* (Cistaceae). *Ann. Bot.* 90, 345–351. doi: 10.1093/aob/mcf194

Felix, W. J. P., Felix, L. P., Melo, N. F., Oliveira, M. B. M., Dutilh, J. H. A., and Carvalho, R. (2011). Karyotype variability in species of the genus *Zephyranthes* Herb.(Amaryllidaceae–Hippeastreae). *Plant Syst. Evol.* 294, 263–271. doi: 10.1007/s00606-011-0467-6

Fereidounfar, S., Ghahremaninejad, F., and Khajehpiri, M. (2016). Phylogeny of the Southwest Asian *Pimpinella* and related genera based on nuclear and plastid sequences. *Genet. Mol. Res.* 15:gmr15048767. doi: 10.4238/gmr15048767

Fregonezi, J. N., Rocha, C., Torezan, J. M. D., and Vanzela, A. L. L. (2004). The occurrence of different Bs in *Cestrum intermedium* and *C. strigilatum* (Solanaceae) evidenced by chromosome banding. *Cytogenet. Genome Res.* 106, 184–188. doi: 10.1159/000079285

Galland, N. (1988). *Recherche sur L'origine de la Flore Orophile du Maroc. Etude Caryologique et Cytogeographique. Travaux de l'Institut Scientifique. Seirie Botanique, No. 35*. Rabat: Universiteì Mohammed V, Institut scientifique, 1–168.

Gawlowska, M. (1967). *Pimpinella nigra* Willd. Poland part III. Numbers of chromosomes in *Pimpinella nigra* Willd. and related species. *Diss. Pharm.* 19, 439–450.

Greilhuber, J., Doležel, J., Lysák, M. A., and Bennett, M. D. (2005). The origin, evolution and proposed stabilization of the terms 'genome size'and 'C-value'to describe nuclear DNA contents. *Ann. Bot.* 95, 255–260. doi: 10.1093/aob/mci019

Grime, J. P., and Mowforth, M. A. (1982). Variation in genome size-an ecological interpretation. *Nature* 299, 151–153. doi: 10.1038/299151a0

Guerra, M. (2008). Chromosome numbers in plant cytotaxonomy: concepts and implications. *Cytogenet. Genome Res.* 120, 339–350. doi: 10.1159/000121083

Hamidi, F., Karimzadeh, G., Monfared, S. R., and Salehi, M. (2018). Assessment of Iranian endemic *Artemisia khorassanica*: karyological, genome size, and gene expressions involved in artemisinin production. *Turk. J. Biol.* 42, 329–340. doi: 10.3906/biy-1802-86

Hesamzadeh Hejazi, S. M., and Ziaei Nasab, M. Z. (2010). Cytotaxonomy of some *Onobrychis* (Fabaceae) species and populations in Iran. *Caryologia* 63, 18–31. doi: 10.1080/00087114.2010.589705

Huziwara, Y. (1962). Karyotype analysis in some genera of Compositae. VIII. Further studies on the chromosomes of Aster. *Am. J. Bot.* 49, 116–119. doi: 10.1002/j.1537-2197.1962.tb14916.x

Jones, N. (2012). B chromosomes in plants. *Plant Biosyst.* 146, 727–737. doi: 10.1080/11263504.2012.713406

Jurtseva, O. V. (1988). The cytologic study of some species of the genus *Pimpinella* L. (Umbelliferae-Apioideae). *Biol. Nauki* 11, 78–84.

Karimzadeh, G., Danesh-Gilevaei, M., and Aghaalikhani, M. (2011). Karyotypic and nuclear DNA variations in *Lathyrus sativus* (Fabaceae). *Caryologia* 64, 42–54. doi: 10.1080/00087114.2011.10589763

Kellogg, E. A., and Bennetzen, J. L. (2004). The evolution of nuclear genome structure in seed plants. *Am. J. Bot.* 91, 1709–1725. doi: 10.3732/ajb.91.10.1709

Knight, C. A., Molinari, N. A., and Petrov, D. A. (2005). The large genome constraint hypothesis: evolution, ecology and phenotype. *Ann. Bot.* 95, 177–190. doi: 10.1093/aob/mci011

Lavania, U. C., and Srivastava, S. (1999). Quantitative delineation of karyotype variation in *Papaver* as a measure of phylogenetic differentiation and origin. *Curr. Sci.* 77, 429–435.

Leitch, I. J., and Bennett, M. D. (2004). Genome downsizing in polyploid plants. *Biol. J. Linn. Soc.* 82, 651–663. doi: 10.1111/j.1095-8312.2004.00349.x

Levan, A., Fredga, K., and Sandberg, A. A. (1964). Nomenclature for centromeric position on chromosomes. *Hereditas* 52, 201–220.

Loureiro, J., Rodriguez, E., Doležel, J., and Santos, C. (2007). Two new nuclear isolation buffers for plant DNA flow cytometry: a test with 37 species. *Ann. Bot.* 100, 875–888. doi: 10.1093/aob/mcm152

Mahdavi, S., and Karimzadeh, G. (2010). Karyological and nuclear DNA content variation in some Iranian endemic *Thymus* species (Lamiaceae). *J. Agric. Sci. Technol.* 12, 447–458.

Maletskii, S. I. (2004). Epigenetic and synergistic types of inheritance of the reproductive characters in angiosperms. *Zh. Obshch. Biol.* 65, 116–135.

Minitab 17 (2010). *Computer Software*. State College, PA: Minitab, Inc.

Mirzaghaderi, G., Karimzadeh, G., Hassani, H. S., Jalali-Javaran, M., and Baghizadeh, A. (2010). Cytogenetic analysis of hybrids derived from wheat and *Tritipyrum* using conventional staining and genomic *in situ* hybridization. *Biol. Plant.* 54, 252–258. doi: 10.1007/s10535-010-0044-9

Mozaffarian, V. (2003). New species and new records of Iranian Umbelliferae. *Bot. J.* 88, 104–124.

Murray, B. G., De Lange, P. J., and Ferguson, A. R. (2005). Nuclear DNA variation, chromosome numbers and polyploidy in the endemic and indigenous grass flora of New Zealand. *Ann. Bot.* 96, 1293–1305. doi: 10.1093/aob/mci281

Navarro, F. B., Suarez-Santiago, V. N., and Blanca, G. (2004). A new species of *Haplophyllum* A. Juss. (Rutaceae) from the Iberian Peninsula: evidence from morphological, karyological and molecular analyses. *Ann. Bot.* 94, 571–582. doi: 10.1093/aob/mch176

Ozkan, H., Tuna, M., and Arumuganathan, K. (2003). Nonadditive changes in genome size during allopolyploidization in the wheat (*Aegilops-Triticum*) group. *J. Hered.* 94, 260–264. doi: 10.1093/jhered/esg053

Palestis, B. G., Trivers, R., Burt, A., and Jones, R. N. (2004). The distribution of B chromosomes across species. *Cytogenet. Genome Res.* 106, 151–158. doi: 10.1159/000079281

Paszko, B. (2006). A critical review and a new proposal of karyotype asymmetry indices. *Plant Syst. Evol.* 258, 39–48. doi: 10.1007/s00606-005-0389-2

Pellicer, J., Garcia, S., Garnatje, T., Dariimaa, S., Korobkov, A. A., and Vallès, J. (2007). Chromosome numbers in some *Artemisia* (Asteraceae, Anthemideae) species and genome size variation in its subgenus *Dracunculus*: karyological, systematic and phylogenetic implications. *Chromosome Bot.* 2, 45–53. doi: 10.3199/iscb.2.45

Pimenov, M. G., Dauschkevich, J. V., Vasil'eva, M. G., and Kljuykov, E. V. (1996). Mediterranean chromosome number reports 6. *Fl. Medit.* 6, 288–307.

Pimenov, M. G., and Leonov, M. V. (1993). *The Genera of the Umbelliferae: A Nomenclator*. Kew: Royal Botanic Gardens.

Pimenov, M. G., Vasil'eva, M. G., Leonov, M. V., and Daushkevich, J. V. (2003). *Karyotaxonomical Analysis in the Umbelliferae*. Enfield, NH: Science Publishers, 57–68,316–326,362–363.

Pu, F. T., and Watson, M. F. (2005). "*Pimpinella* L," in *Flora of China*, eds Z. Y. Wu and P. H. Raven (St. Louis, MO: Miss Bot Garden Press), 93–104.

Romero-Zarco, C. (1986). A new method for estimating karyotype asymmetry. *Taxon* 35, 526–530. doi: 10.2307/1221906

SAS (2003). *SAS 9.1 (version SAS 9.1. 3, Service Pack 3)*. Cary, NC: SAS Institute Inc.

Scheid, O. M., Jakovleva, L., Afsar, K., Maluszynska, J., and Paszkowski, J. (1996). A change of ploidy can modify epigenetic silencing. *Proc. Natl. Acad. Sci. U.S.A.* 93, 7114–7119. doi: 10.1073/pnas.93.14.7114

Seijo, J. G., and Fernández, A. (2003). Karyotype analysis and chromosome evolution in South American species of *Lathyrus* (Leguminosae). *Am. J. Bot.* 90, 980–987. doi: 10.3732/ajb.90.7.980

Shner, J. V., Pimenov, M. G., Kljuykov, E. V., Alexeeva, T. V., Ghahremaninejad, F., and Mozaffarian, V. (2004). Chromosome numbers in the Iranian Umbelliferae. *Chromosome Sci.* 8, 1–9.

Soltis, D. E., Soltis, P. S., Bennett, M. D., and Leitch, I. J. (2003). Evolution of genome size in the angiosperms. *Am. J. Bot.* 90, 1596–1603. doi: 10.3732/ajb.90.11.1596

Srivastava, M. S. (2002). *Methods of Multivariate Statistics.* New York, NY: John Wiley and Sons, 697.

Stebbins, G. L. (1971). *Chromosome Evolution in Higher Plants*. London: Edward Arnold Press.

Suda, J., Kyncl, T., and Freiová, R. (2003). Nuclear DNA amounts in Macaronesian angiosperms. *Ann. Bot.* 92, 153–164. doi: 10.1093/aob/mcg104

Temsch, E. M., Temsch, W., Ehrendorfer-Schratt, L., and Greilhuber, J. (2010). Heavy metal pollution, selection, and genome size: the species of the Žerjav study revisited with flow cytometry. *J. Bot.* 2010:596542. doi: 10.1155/2010/596542

Tuna, M., Vogel, K. P., Arumuganathan, K., and Gill, K. S. (2001). DNA content and ploidy determination of bromegrass germplasm accessions by flow cytometry. *Crop Sci.* 41, 1629–1634. doi: 10.2135/cropsci2001.4151629x

Verlaque, R., and Filosa, D. (1992). Mediterranean chromosome number reports 2 (107–117). *Fl. Medit.* 2, 264–272.

Wolff, H. (1927). Umbelliferae-Apioideae-Ammineae-Carinae, Ammineae novemjugat. et genuinae. (Cryptotaeniopsis). *Pflanzenr* 90, 174–182. doi: 10.1007/s004120000074

Wyman, J., Nicole, B., Denis, F., and Sylvie, L. (1992). Ploidy level stability of callus tissue, axillary and adventitious shoots of *Larix× eurolepis* Henry regenerated *in vitro*. *Plant Sci.* 85, 189–196. doi: 10.1016/0168-9452(92)90115-3

Yurtseva, O. V. (1988). The cytologic study of some species of the genus *Pimpinella* L. (Umbelliferae–Apioideae). *Biol. Nauki* 11, 78–85.

Yurtseva, O. V., and Tikhomirov, V. N. (1998). Morphological diversity and taxonomy of the *Pimpinella tragium* VILL. group (Umbelliferae-Apioideae) in the Mediterranean. *Feddes Repert.* 109, 479–500. doi: 10.1002/fedr.19981090703

Zhou, J., Gong, X., Downie, S. R., and Peng, H. (2009). Towards a more robust molecular phylogeny of Chinese Apiaceae subfamily Apioideae: additional evidence from nrDNA ITS and cpDNA intron (*rpl16* and *rps16*) sequences. *Mol. Phylogenet. Evol.* 53, 56–68. doi: 10.1016/j.ympev.2009.05.029

Zhou, J., Peng, H., Downie, S. R., Liu, Z. W., and Gong, X. (2008). A molecular phylogeny of Chinese Apiaceae subfamily Apioideae inferred from nuclear ribosomal DNA internal transcribed spacer sequences. *Taxon* 57, 402–416. doi: 10.2307/25066012

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

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

Copyright © 2022 Mehravi, Ranjbar, Najafi-Zarrini, Mirzaghaderi, Hanifei, Severn-Ellis, Edwards and Batley. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 27 June 2022
doi: 10.3389/fpls.2022.881879

The Tracking of Moist Habitats Allowed *Aiphanes* (Arecaceae) to Cover the Elevation Gradient of the Northern Andes

María José Sanín[1,2,3], Finn Borchsenius[4], Margot Paris[5], Sara Carvalho-Madrigal[1], Andrés Camilo Gómez Hoyos[1], Agustín Cardona[3], Natalia Arcila Marín[1], Yerson Ospina[1], Saúl E. Hoyos-Gómez[6], Héctor Favio Manrique[7] and Rodrigo Bernal[8]*

[1] Facultad de Ciencias y Biotecnología, Universidad CES, Medellín, Colombia, [2] School of Mathematical and Natural Sciences, Arizona State University, Tempe, AZ, United States, [3] Departamento de Procesos y Energía, Universidad Nacional de Colombia, Medellín, Colombia, [4] Faculty of Technical Sciences, Aarhus University, Aarhus, Denmark, [5] Unit of Ecology and Evolution, Department of Biology, University of Fribourg, Fribourg, Switzerland, [6] Instituto de Biología, Universidad de Antioquia, Medellín, Colombia, [7] Jardín Botánico del Quindío, Armenia, Colombia, [8] Reserva Natural Guadualito, Montenegro, Colombia

OPEN ACCESS

Edited by:
Gerald Matthias Schneeweiss, University of Vienna, Austria

Reviewed by:
Tim Böhnert, University of Bonn, Germany
Gwendolyn Peyre, University of the Andes, Colombia

***Correspondence:**
María José Sanín
msanin2@asu.edu

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

Received: *23 February 2022*
Accepted: *20 May 2022*
Published: *27 June 2022*

Citation:
Sanín MJ, Borchsenius F, Paris M, Carvalho-Madrigal S, Gómez Hoyos AC, Cardona A, Arcila Marín N, Ospina Y, Hoyos-Gómez SE, Manrique HF and Bernal R (2022) The Tracking of Moist Habitats Allowed Aiphanes (Arecaceae) to Cover the Elevation Gradient of the Northern Andes. Front. Plant Sci. 13:881879. doi: 10.3389/fpls.2022.881879

The topographic gradients of the Tropical Andes may have triggered species divergence by different mechanisms. Topography separates species' geographical ranges and offers climatic heterogeneity, which could potentially foster local adaptation to specific climatic conditions and result in narrowly distributed endemic species. Such a pattern is found in the Andean centered palm genus *Aiphanes*. To test the extent to which geographic barriers and climatic heterogeneity can explain distribution patterns in *Aiphanes*, we sampled 34 out of 36 currently recognized species in that genus and sequenced them by Sanger sequencing and/or sequence target capture sequencing. We generated Bayesian, likelihood, and species-tree phylogenies, with which we explored climatic trait evolution from current climatic occupation. We also estimated species distribution models to test the relative roles of geographical and climatic divergence in their evolution. We found that *Aiphanes* originated in the Miocene in Andean environments and possibly in mid-elevation habitats. Diversification is related to the occupation of the adjacent high and low elevation habitats tracking high annual precipitation and low precipitation seasonality (moist habitats). Different species in different clades repeatedly occupy all the different temperatures offered by the elevation gradient from 0 to 3,000 m in different geographically isolated areas. A pattern of conserved adaptation to moist environments is consistent among the clades. Our results stress the evolutionary roles of niche truncation of wide thermal tolerance by physical range fragmentation, coupled with water-related niche conservatism, to colonize the topographic gradient.

Keywords: climatic, environmental niche, geographical overlap, narrow endemic, palms, realized niche, species distribution models, phylogenomics

INTRODUCTION

The Tropical Andes Biodiversity hotspot, also referred to as the uplands of Western Amazonia, spans from Venezuela, Colombia, Ecuador, Peru, Bolivia to Northern Argentina (Myers et al., 2000; Mittermeier et al., 2004, 2011). It ranks first among 36 world hotspots for biodiversity based on species richness and endemism and level of threat, and is estimated to contain nearly one-sixth of all vascular plant species. The causal mechanisms behind the explosion of species richness during the ongoing orogeny of the Tropical Andes have been extensively discussed in the last 2 decades (Hoorn et al., 2010; Antonelli and Sanmartín, 2011; Luebert and Weigend, 2014; Antonelli et al., 2018; Rahbek et al., 2019), and several factors associated with mountain building were identified to promote the extraordinary taxonomic diversification in these areas (Graham et al., 2014; Benham and Witt, 2016; Antonelli et al., 2018). With the creation of a remarkable diversity of novel heterogeneous habitats, organisms could adapt to and/or specialize in new topographic complexity and climatic conditions (i.e., temperature and orographic precipitation). In addition to the strong climatic and ecological gradients that characterize mountain areas, uplift and erosion form new physically constrained habitats potentially leading to a high proportion of mountain endemics by allopatric isolation (Antonelli and Sanmartín, 2011). A particular challenge in determining which factors promote diversification comes from the intrinsic relationship between topographic and climatic gradients during mountain evolution. Furthermore, both space and time dynamics of these factors are important such as climatically driven connection and disconnection of populations (Flantua et al., 2019), speciation extinction and migration over macroevolutionary time (Chazot et al., 2016, 2018), geodiversity (Muellner-Riehl et al., 2019; Rahbek et al., 2019), age, and isolation (Rahbek et al., 2019).

Andean taxa can be largely composed of rare and narrow endemics that require substantial local sampling in hardly accessible sites. Our knowledge of this diversity in the Northern Andes has grown over the last years, especially in Colombia where many areas of the country have now become available for research. This has made it possible to use the Andean palm genus *Aiphanes* Willd. (Arecoideae: Cocoseae: Bactridinae) as a test case. These palms are commonly narrow mountain endemics (Bernal and Borchsenius, 2010; Bernal et al., 2019a), even more so than previously thought, with the description of 14 species in the last 3 decades since the latest monograph (Borchsenius and Bernal, 1996), of which 6 species described in the last 5 years are known from only one locality (Bernal et al., 2017, 2019a,b).

Aiphanes includes 36 species restricted to the understory of lowland or montane forests of the Neotropics, with most species between 6° N and 4° S in the Andes or in the surrounding areas of Western Amazonia and the Choco (**Figure 1A**). It also includes species that grow up to 21 m tall, like *A. pilaris* R. Bernal and *A. grandis* Borchs. & Balslev, but most are medium-sized, understory palms, and some of them very small or acaulescent. A couple of species grows in dry seasonal habitats (*A. eggersii* Burret *and A. horrida* Burret), one at the mountain tree line (*A. verrucosa* Borchs. & Balslev), and several in the Andean foothills (i.e., *A. macroloba* Burret, *A. acaulis* Galeano & R. Bernal, and *A. buenaventurae* R. Bernal & Borchs), and a single species (*A. argos* R. Bernal, Borchs. & Hoyos-Gómez) is restricted to riverine habitats. Most species occur as inconspicuous or rare elements in cloud forests like *Aiphanes verrucosa* in Ecuador (Svenning et al., 2009).

Definitely, one of the most striking characteristics of *Aiphanes,* when compared to other species-rich Neotropical palm genera, is its paucity. It is uncommon to see extensive or dense populations of *Aiphanes* as it is to see several species of this genus coexisting at the same site. This contrasts with other speciose Neotropical forest palm genera like *Geonoma* Willd. and *Chamaedorea* Willd., which frequently form prominent and abundant populations and local assemblages of several congenerics dominating the understory. Despite their paucity, *Aiphanes* palms have conquered the elevation gradient of the Andes (which we here understand as the foothills and montane environments between 0 and 3,000 m above sea level), whereas most other Neotropical palm genera have not. Do geographical or climatic factors determine this genus' success in conquering the gradient despite its paucity, low local diversity, and high endemicity?

Moist tropical climates have favored diversification in palms (Svenning et al., 2008), which could be due to higher population sizes (which appear unlikely in *Aiphanes*), biotic interactions, greater ecological success of palms given their specific morphology and anatomy, or water-energy dynamics (Eiserhardt et al., 2011a). Other studies on the geographical ecology of palms suggest that precipitation seasonality can be the most important climatic predictor of species richness (in western Amazonia: Kristiansen et al., 2011), especially if combined with temperature seasonality extremes (in Brazil: Salm et al., 2007). In South America, Neves et al. (2020) found that closely related lineages likely phylogenetically conserve adaptation to a precipitation regime. Here, we want to evaluate the role of geography and climate despite the fact that biotic interactions, morphology, and population sizes can also play important roles.

Phylogenetic studies provide the backbone for hypothesis testing of a causal mechanism underlying divergence. In the case of climatic specialization, they can be used to compare resulting niche breadths of sister species (Bonetti and Wiens, 2014). Different morphological and ecological traits can be compared between same-clade species (i.e., species belonging to a monophyletic clade in the genus) and across the evolution of taxa from different clades to test whether these differ more or less in any groups in particular (Graham et al., 2004; Schnitzler et al., 2012). This, used in combination with species distribution models (SDMs), allows for exploration of the role of geography and climate in species divergence. In principle, phylogenetic niche conservatism states that phylogenetically related species should share the tendency to occupy similar (climatic) niches (Harvey and Pagel, 1991). Similarly, closely related species should occupy closer geographical areas than distantly related ones or should have largely geographically overlapping distributions unless geography is the factor driving their divergence.

Furthermore, dispersal constraints caused by physical barriers can truncate species' fundamental climatic niche (Feeley and Silman, 2010), enforcing differentiation of the realized niche

between closely related species. In Australia, thermal truncation was prevalent especially in forest species. In that study, the truncation of water-related climatic space was not assessed (Bush et al., 2018). In the case of mountains, where physical barriers are known to be important factors causing species divergence (Antonelli et al., 2018), we expect the realized climatic niche of species to become a specialized subset of their potential climatic niche resulting in small realized niche widths. We also expect that the capacity to specialize (from a wider fundamental climatic tolerance) plays a role in species' colonization of the elevation gradient. Here, we seek to understand which aspects of the climatic niche are conserved and which are free to change and specialize in the midst of physical range fragmentation.

In this article, we assess two main drivers for *Aiphanes* divergence patterns that correlate to mountain building across the Andean Cordilleras. The first one alludes to climatic specialization, which was assessed by current climatic space occupancy based on herbarium records, and its evolution was reconstructed on phylogenies. The second one alludes to physical isolation of populations over time (divergence mediated by geographical isolation) assessed by SDMs and their overlap between species. In contrast to the expectations of phylogenetic niche conservatism, we predict same-clade species to occupy a significantly different (non-overlapping) climatic space. We also predict non-overlapping geographical distributions between same-clade species. Thus, we expect species to depart from climatic niche conservatism because topographic gradients offer steep and wide climatic gradients in which species can specialize favored by isolation in physically confined areas. These areas truncate species climatic niche. We conducted an extensive field collection for genetic sampling covering all but two species of *Aiphanes* (*A. bio* R. Bernal, Borchs., Hoyos-Gómez, H.F. Manrique & Sanín and *A. cogollo* R. Bernal, Borchs., Hoyos-Gómez, H.F. Manrique & Sanín) to build genus-wide phylogenies combining classical Sanger sequencing with a target enrichment bait set of more than 4,000 targeted genes. Phylogenetic relationships are used to discuss how the climatic and geographical occupations contrast between same-clade species, and how climate-related traits evolve between clades and species.

MATERIALS AND METHODS

Taxon Sampling and Genomic Data Obtainment

DNA Extraction

For all samples in both the target capture and Sanger sequencing datasets, we used a DNeasy® plant mini kit (Qiagen, Venlo, Netherlands) following the supplier's instructions for DNA isolation. DNA quantities were measured with a NanoDrop™ spectrophotometer (Thermo Fisher Scientific, Waltham, MA, United States). All extracted samples and their use for different phylogenies are listed in **Supplementary Tables 1**, **2**.

Sampling for the Sanger Sequencing Phylogeny

We used 64 samples (**Figure 1B**) representing 31 of the 36 species currently accepted in the genus. Whenever possible, we included several individuals per species; for widely distributed species with described subspecies or forms (such as in *A. hirsuta* where there are four described subspecies: Borchsenius and Bernal, 1996), we aimed at sampling from geographically distant localities. Furthermore, we sampled 11 individuals representing 11 species of the other four genera of the subtribe Bactridinae (*Acrocomia* Mart., *Astrocaryum* G. Mey., *Bactris* Jacq. ex Scop., and *Desmoncus* Mart.).

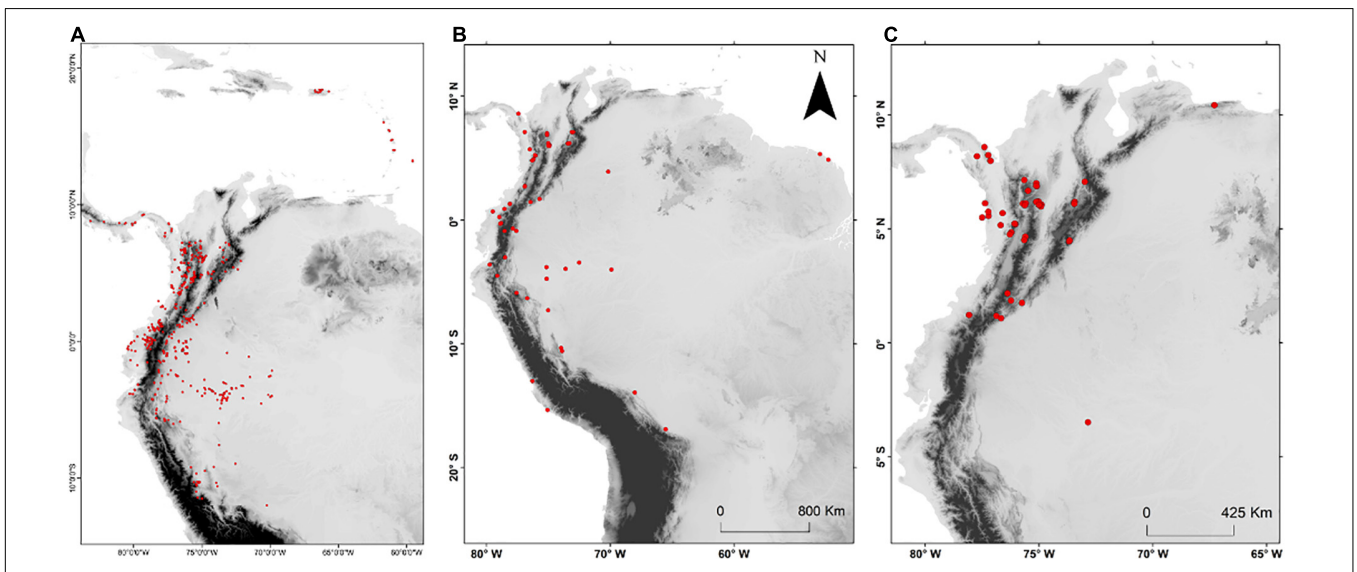

FIGURE 1 | Curated occurrence records for **(A)** all species of *Aiphanes* used in this study, **(B)** all samples used for Sanger sequencing, and **(C)** all samples used for sequence capture.

Amplification and Sequence Preparation

For Sanger sequencing, we amplified three nuclear genome regions that are commonly used in palm systematics: ITS, prk, and rpb2. We used classical PCR parameters, as described by Eiserhardt et al. (2011b). Chromatograms were visually checked for erroneous or ambiguous nucleotide calling. **Supplementary Table 1** lists all the samples included and genes that were amplified for each sample, as well as the accession codes for each marker on GenBank. DNA Sequences were aligned for each gene independently using Muscle (Edgar, 2004) as implemented in the EMBL-EBI search and sequence analysis tools (Madeira et al., 2019) followed by a visual check.

Sampling for the Sequence Capture Phylogeny

We gathered 98 samples (**Figure 1C**) representing 23 species of predominantly Colombian species complexes of *Aiphanes,* totaling 64% of the genus and focusing on groups that we considered problematic or poorly studied including *Aiphanes parvifolia* Burret and recently described and morphologically similar species (Bernal et al., 2019a), *A. lindeniana* H. Wendl., *A. hirsuta* Burret, and *A. simplex* Burret complexes. In these cases and in few others, we included 2–31 individuals per species, covering different forms or subspecies that have been described as well as the widest possible geographic distribution. The most densely sampled species was *A. hirsuta*, covering all four recognized subspecies. The sampling concentrated on the Andes of Colombia, on the Central, Western Cordilleras, and the Pacific lowlands (including Gulf of Tribugá where it had not been previously registered), where taxonomic novelties and micro-endemism seemed more relevant because of recent findings (Bernal et al., 2019a,b). We also sampled 24 individuals in 23 species from other palm genera as outgroups (refer to **Supplementary Table 2** for a full list with coordinates, herbarium samples, and accession codes). For a brief description of the sampling schemes for both sequencing approaches, refer to **Table 1** and **Supplementary Material** in this article.

Dual-Indexed Library Preparation and Target Capture Sequencing

For nuclear target sequencing, a total of 500 ng was fragmented to 400-bp fragments with a Bioruptor® ultrasonicator (Diagenode, Liège, Belgium). Library preparations were performed following de La Harpe et al. (2019) and using a KAPA LTP library preparation kit (Roche, Basel, Switzerland) for sample cleaning, end-repair, and A-tailing steps, and the protocol of Meyer and Kircher (2010) for adaptor ligation and adaptor fill-in reactions steps. Four μl of the ligated fragment solution were amplified for eight cycles using KAPA HiFi DNA Polymerase (Roche, Basel, Switzerland) and the set of 60 dual index primers described in Loiseau et al. (2019). Libraries were quantified with a Qubit® Fluorometer v2.2 before pooling in equimolar ratio. Target capture was performed using the custom kit PopcornPalm developed by de La Harpe et al. (2019), and targeting 4,051 palm genes. Target capture was conducted on pools of 50 or 51 samples, following myBait® Custom Target Capture Kits protocol v3.0 (Arbor Biosciences, Ann Arbor, MI, United States), with 18 h of incubation time at 65°C and 12 cycles of post-capture PCR reactions. The pooled target capture reactions were quantified with Qubit® Fluorometer v 2.2, before sequencing with an Illumina HiSeq 3000 sequencer (Illumina, San Diego, CA, United States) in paired-end 2 × 150-bp mode.

TABLE 1 | Number of *Aiphanes* individuals sampled for each sequencing approach.

	Species	Sanger sequence	Sequence capture
1	*Aiphanes acaulis* Galeano & R. Bernal	1	1
2	*Aiphanes argos* R. Bernal, Borchs., Hoyos-Gómez	2	2
3	*Aiphanes bicornis* Cerón & R. Bernal	1	0
4	*Aiphanes bio* R. Bernal, Borchs., Hoyos-Gómez, H.F. Manrique & Sanín	0	0
5	*Aiphanes buenaventurae* R. Bernal & Borchs.	1	2
6	*Aiphanes chiribogensis* Borchs. & Balslev	1	0
7	*Aiphanes cogollo* R. Bernal, Borchs., Hoyos-Gómez, H.F. Manrique & Sanín	0	0
8	*Aiphanes concinna* H.E. Moore	0	7
9	*Aiphanes decipiens* R. Bernal, Borchs., Hoyos-Gómez, H.F. Manrique & Sanín	1	2
10	*Aiphanes deltoidea* Burret	1	1
11	*Aiphanes duquei* Burret	1	0
12	*Aiphanes eggersii* Burret	2	0
13	*Aiphanes erinacea* (H. Karst.) H. Wendl.	4	1
14	*Aiphanes gelatinosa* H.E. Moore	0	1
15	*Aiphanes gloria* R. Bernal, Borchs., Hoyos-Gómez, H.F. Manrique & Sanín	2	3
16	*Aiphanes graminifolia* Galeano & R. Bernal	1	0
17	*Aiphanes grandis* Borchs. & Balslev	1	0
18	*Aiphanes hirsuta* Burret	6	31
19	*Aiphanes horrida* (Jacq.) Burret	6	3
20	*Aiphanes killipii* (Burret) Burret	1	2
21	*Aiphanes leiostachys* Burret	1	2
22	*Aiphanes lindeniana* (H. Wendl.) H. Wendl.	3	9
23	*Aiphanes linearis* Burret	2	12
24	*Aiphanes macroloba* Burret	1	3
25	*Aiphanes minima* (Gaertn.) Burret	1	0
26	*Aiphanes multiplex* R. Bernal & Borchs.	1	0
27	*Aiphanes parvifolia* Burret	1	2
28	*Aiphanes pilaris* R. Bernal	1	1
29	*Aiphanes simplex* Burret	2	5
30	*Aiphanes spicata* Borchs. & R. Bernal	2	0
31	*Aiphanes suaita* R. Bernal, Sanín & Castaño	1	3
32	*Aiphanes tatama* R. Bernal, Borchs., Hoyos-Gómez, H.F. Manrique & Sanín	1	1
33	*Aiphanes tricuspidata* Borchs., R. Bernal & M. Ruiz	1	2
34	*Aiphanes ulei* (Dammer) Burret	7	2
35	*Aiphanes verrucosa* Borchs. & Balslev	1	0
36	*Aiphanes weberbaueri* Burret	6	0
	Assorted Arecaceae	11	24
	TOTAL	75	122

Read Trimming, Mapping, and SNP Calling

The program ConDeTri v2.2 (Smeds and Künstner, 2011) was used to trim the raw reads, with 20 as high-quality threshold parameter. The program bowtie2 v2.2.5 (Langmead and Salzberg, 2012) with the very sensitive local option was used for read mapping. We used the *Geonoma undata* reference genome (de La Harpe et al., 2019) for mapping; it was the closest reference genome available. The proportion of in−target reads (specificity) and the proportion of baits covered (efficiency) were calculated for each sample using bedtools v2.24.0 (Quinlan and Hall, 2010) following de La Harpe et al. (2019). The target capture method was highly successful for all the *Aiphanes* species and samples included in our analyses, with average specificity of 79.3% (range: 74.9–81.1%) and average efficiency of 91.3% (range: 76.4–95.4%). We then selected reads mapping at a unique location on the genome and masked PCR duplicates with Picard tools v1.119[1]. The program GATK v3.8 (McKenna et al., 2010) was used to realign the reads around indels, base-recalibration, and SNPs calling for target regions, using UnifiedGenotyper with the EMIT_ALL_SITES option in order to obtain both variable and invariable sites. Sites were filtered using VCFtools v0.1.13 (Danecek et al., 2011), with a minimum quality of 20, a minimum depth of 8× per sample, and a maximum of 50% of missing data, and by removing indels. After filtering, the targeted capture method provided a total of 2.557.512 high-quality sequenced bases with an average coverage of 29.2× per sample and distributed in 2,867 genes. The bait kit developed by de La Harpe et al. (2019) for micro- and macro-evolutionary analyses of palms is large (4,051 genes) and contains fast- to slow-evolving DNA regions. We included the whole set of markers for this study, because we sampled both at the species and "morphotype" levels. We also aimed at having a robust phylogeny that relies on informative data and could serve as a backbone or a phylogenetic framework for future studies involving more specific questions on *Aiphanes*.

Phylogenetic Inference

Phylogenetic Analysis of the Sanger Sequences

Two Bayesian phylogenies were generated in BEAST v 1.10 (Suchard et al., 2018). The three different DNA regions were concatenated and partitioned, and the site model was unlinked. The evolutionary site model was selected using the Akaike Information Criterion (GTR + *invariant sites* for ITS, TN92 for prk, and GTR+GAMMA for rpb2) in jModelTest (Guindon and Gascuel, 2003; Darriba et al., 2012). We chose a lognormal uncorrelated relaxed clock model to account for rate heterogeneity and the birth death tree branching prior. The analyses were run for 2 chains of 100.000 generations each. ESS values (>200) and chain convergence were assessed in TRACER v 1.7.1 (Rambaut et al., 2018). For each different analysis, trees were combined and summarized in the LogCombiner and TreeAnnotator (maximum clade credibility tree with a posterior probability limit of 0.5 and burn-in of 10%) applications of the BEAST 1.10 package (Suchard et al., 2018). The resulting maximum clade credibility tree will be hereafter referred to as the Sanger sequence phylogeny (SSP).

Phylogenetic Analyses of the Target Capture Sequences

Phylogenetic trees were estimated using the maximum-likelihood and coalescent-based species tree methods. The choice to use both methods stems from the expectation that different gene trees could lead to different phylogenetic relationships (Liu et al., 2015a), something that could not be accounted for using concatenated DNA regions in a Maximum Likelihood reconstruction alone. This was of particular importance in this data set where we concentrated our sampling in species complexes that included several described forms or newly described taxa. It was our priority to be able to conduct species clade assignment in the downstream analyses that relies on accounting for possible incongruence between gene tree topologies. The concatenated alignment of the 2,557.512 high-quality bases, including both variable and invariable sites and distributed in 2867 genes, was analyzed with RAxML v8.228 (Stamatakis, 2014) using the GTR+GAMMA model of substitution and 100 rapid bootstrap replicates. The concatenation of a large number of genes often results in phylogenetic trees with high node support values, but the assumption that all genes share the same topology and branch lengths is often violated and can lead to high support for the wrong topology (Kubatko and Degnan, 2007). Coalescent-based methods are better suited for datasets with multiple loci, as they consider gene tree incongruence due, for example, to incomplete lineage sorting (Liu et al., 2009). We therefore used ASTRAL v5.6.1 with default parameters; ASTRAL is a faster, two-steps coalescent-based method that estimates the species tree, given a set of gene trees (Mirarab et al., 2014; Mirarab and Warnow, 2015). We used the *-a* option from Rabiee et al. (2019), 515 genes and 40 bootstrap replicates per gene, with each gene contributing equally to localPP support. For gene selection, a first list of genes was selected based on missing data, retaining 1,993 genes with more than 400-bp sequence length covered after filtering for high-quality bases, including both variable and invariable sites, and with sequence information for all the samples (i.e., no sample consisted entirely of missing data). Gene trees were first estimated with RAxML v8.228 (Stamatakis, 2014) using the GTR+GAMMA model of substitution and 100 bootstrap replicates. Weakly informative gene trees with average bootstrap values lower than 40 were not kept for further ASTRAL analysis in order to avoid a potential decrease in the accuracy of species tree estimation (Liu et al., 2015b; Molloy and Warnow, 2018). A total of 515 highly informative selected genes exhibited an average length of 1,894 bp (range: 475–11,383 bp), more than two times higher than the average length of 920 bp for weakly informative genes (range: 400–5,112 bp). This number of highly informative genes detected with our analyses involving mainly *Aiphanes* samples is concordant with the number of 795 highly informative genes detected by Loiseau et al. (2019) using the same capture kit and 20 palm samples representing a wide range of evolutionary time scales, from intra-specific variability of up to

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

88 Ma of divergence. The RAxML and the ASTRAL trees will be hereafter referred to as the sequence capture phylogenies (S).

Phylogenetic Dating

The SSP was dated by the following secondary calibrations obtained from Eiserhardt et al. (2011b): core north Andean clade at 11 My and crown of genus at 28 My, both using normal distribution with a standard deviation of 1, a log-normal relaxed clock, and a birth death branching prior.

Dating of the SCPs was performed by penalized likelihood using the function *chronos* in "ape" R package v. 5.4–1 (Paradis et al., 2004; Paradis, 2013) (lambda smoothing parameter = 2, model = correlated) and node calibration for the crown *Aiphanes* (age.min = 27 and age.max = 29), and a second calibration age resulting from population parameter estimation in the coalescent. This divergence age was estimated for *lindeniana* + *linearis* clades in SNAPP (Bryant et al., 2012) implemented in Beast 2.6 (Bouckaert et al., 2019) on CIPRES Science Gateway version 3.3 (Stamatakis et al., 2008; Miller et al., 2010). For this age estimation, we included all sampled individuals of *A. hirsuta, A. lindeniana, A. linearis,* and *A. concinna* in the SCP. We ran 5 chains with a randomly resampled matrix of 800 nuclear SNPs for 10 million generations and then checked them in Tracer v (Rambaut et al., 2018) for chain convergence, sampling efficiency of the priors, likelihood and posterior as well as mutation rates, and theta parameters (these two were left to estimate). Time was obtained from node height of the maximum clade credibility tree after 10% burn-in by conversion using the mutation rate for corn ($m = 2.61 \times 10^{-9}$) (Gaut et al., 1996).

Species Environmental Distribution Models

Species ranges were estimated by extracting species occurrence data from GBIF using the "rgbif" package (version 2.2.0) (Chamberlain et al., 2022) and other data sources (Herbaria: HUA, JAUM, and UTMC). To clean up common errors in the occurrence data, we used QGIS (version 10.2). The coordinates were compared with maps made by experts (Henderson et al., 1995; Borchsenius and Bernal, 1996; Dransfield et al., 2008; Bernal and Borchsenius, 2010; Galeano and Bernal, 2010), as we explain below. Expected distributions were taken from drawn polygons from the existing literature (cited above) and were overlaid to each species coordinates to check for outliers. When we found a dubious point, it was checked by RB (coauthor). If the point could not be verified (or the taxonomy was dubious, meaning the point could have been erroneously assigned to another species), it was discarded.

Georeferencing precision was assessed following Escobar et al. (2016) and verifying with official Base Cartography that the sites described coincided with the coordinates of the biological records. Multicollinearity was evaluated on bioclimatic layers from where the accessible area (m area for species) was cut using an "Extract by mask" algorithm (ArcMap algorithm). Accessible areas (m) were defined by selecting Olson ecoregions that intersected with biological records (Barve et al., 2011). We also performed attribute filtering to eliminate incomplete information, duplicate records, and data from years before 1979. Finally, we stored this information in a geodatabase using the ArcCatalog software (version 10.5).

We used the 19 bioclimatic variables developed by Karger et al. (2017) and available online. We chose CHELSA layers because they incorporate corrections for the effect of wind, the boundary layer, and exposures in mountain valleys (Karger et al., 2017). The layers Bio1-Bio19 incorporate data from the time period of January 1979 to December 2013 and are available at 1 km (30 arcs) resolution. We extracted all climatic layer values for each curated occurrence point using the accessible areas (M) as a mask. This M area was based on the delineation of ecoregions (Olson et al., 2001) and the species range proposed by different authors (Henderson et al., 1995; Galeano and Bernal, 2010; Galeano et al., 2015). We eliminated the correlated predictors using the variance inflation factor (VIF) with the R package *usdm* v 1.18 (Naimi et al., 2014). Seven bioclimatic variables did not exhibit collinearity (VIF < 10) (**Supplementary Table 3**). All used coordinates fell within 20° N and 18° S and 60° W and 81° W.

Species distribution models were generated with Maxent maximum entropy algorithm v 3.4.1 (Phillips et al., 2017). This algorithm has been widely used, generating adequate results in exploration of niches and species distribution (Altamiranda-Saavedra et al., 2017; Calixto-Pérez et al., 2018). Maxent has the advantage of providing an evaluation of omission/commission, response curves, and analysis of variable contributions, which is highly useful for understanding the outputs of the model. The background points were generated on the M area to avoid inflation of the AUC. We generated 10 repetitions per species applying the bootstrapping technique and randomly partitioned the species data in each replicate (85% training and 15% validation). All the models were run using default settings (10,000 background points, 500 maximum iterations with a 10^{-5} convergence threshold, regularization multiplier of 1, and duplicate occurrence removal). The predictive capacity of each model was assessed using the area under the curve (AUC) value that is generated using the ROC (receiver operating characteristic) technique performed by Maxent. The best model was kept ensuring a test AUC ≥ 0.9 for each species. Finally, the best model of each species was projected to geography and reclassified using a threshold that represents the lowest rate of omission in the data training and testing (<15%). The result was a binary polygon (1/0: presence/absence) (Elith et al., 2011; Phillips et al., 2017). For species with few records and known as endemic to only a few pixels, manual models were produced by reclassifying the digital elevation model Alos Palsar (pixel size: 12.5 m × 12.5 m) to the elevation reported in the coordinates of the few herbarium samples available, corresponding to the narrow distribution of the species. Then, the reclassified raster was cut with watersheds that included all the coordinates.

Measuring Niche Overlap

We tested for niche overlap following the environmental-PCA method proposed by Broennimann et al. (2012) and implemented in the R package *ecospat* v. 3.1 (Di Cola et al., 2017). A PCA on the selected predictors by the VIF was computed, and the resulting environmental space for the study area was gridded at

1 km^2 cell resolution. Then, a smooth kernel density function was applied to the occurrence records plotted on the gridded environmental space. The observed niche overlap score for each species pair was estimated with Schoener's *D* metric, which ranges from 0 (no overlap) to 1 (complete overlap) (Warren et al., 2008). Then, statistical tests for niche similarity and niche equivalence hypotheses (Warren et al., 2008) were performed. The first test evaluated whether the ecological niche occupied by two lineages were identical and the second assessed whether the ecological niches of two entities were more or less similar than expected by chance. We repeated each test 100 times, returning a null distribution of overlap values to which the observed niche overlap (D) was compared. If the observed *D* value fell outside of the 95th percentile of expected *D* values, the null hypothesis of random equivalency/similarity was rejected. Finally, we evaluated if the mean *D* values per clade was negatively correlated to the age of each clade [Pearson method implemented in *cor.test* function of the *stats* package in R Core Team (2020)].

Geographic Distribution and Climatic Evolution

The climatic data for statistical analysis were extracted from the previously curated occurrence points overlaid on the CHELSA layers. The temperature-related (Bio1-Bio11) and precipitation-related variables (Bio12-Bio19) were explored by principal component analyses on temperature and precipitation biplots. Four variables were chosen based on their orthogonality in the PCA, on Sanín et al. (2016), and on our knowledge of factors that could be related to topography and that could determine plant growth; these were: mean annual temperature (Bio1), temperature during the coldest month (Bio6), annual precipitation (Bio12), and precipitation seasonality (Bio15). Although we could have reconstructed the seven variables from VIF, this analysis was exploratory and independent from the SDMs; our aim was to see how they evolved on the phylogenies. These variables were reconstructed as continuous traits on the SCP and SSP phylogenies using the *Rphylopars* R package v 0.3.2 (Bruggeman et al., 2009) under the Brownian motion default function of continuous trait evolution by Ho and Ané (2014) and using the mean of each species of all values extracted from all occurrences available for each species. We also used these values to conduct a PGLS (phylogenetic generalized least squares) test for phylogenetic signal of two variable sets: the temperature-related and precipitation-related variables using the *pgls* and *pgls.profile* functions of the package *caper* v 1.0.1 (Orme et al., 2012).

RESULTS

Phylogenetic Reconstructions

The SSP (**Figure 2**) shows seven clades that agree with the SCPs, although these clades are not all well-supported [posterior probabilities (PP) of 1 for clades *acaulis*, *weberbaueri*, and *horrida*; PP of 0.97 for clades *parvifolia* and *linearis*; PP of 0.72 for clade *lindeniana*; PP of 0.54 for clade *simplex*]. Support for this seven-clade backbone is strengthened in the SCPs that we discuss below (for support values in all trees, refer to **Table 2**). Thus, the SSP provides a backbone showing several important relationships: (a) *Aiphanes killipii* as sister to the rest of the genus, (b) the *horrida* clade is supported, with three allopatrically distributed species, (c) *A. grandis* as sister to the mainly Andean species placed by Burret (1932) in subgenus *Brachyanthera*, and (d) the main pattern inside the *Brachyanthera* clade. This main pattern shows a well-supported *weberbaueri* clade, a well-supported *acaulis* clade, and the remaining predominantly Colombian species organized into four clades (*lindeniana*, *linearis*, *parvifolia*, and *simplex* clades).

The SCPs provide additional resolution to groups that the Sanger tree did not resolve. The targeted capture method provided a total of 2,557,512 high-quality sequenced bases, with an average coverage of 29.2× per sample. The Maximum Likelihood concatenated RAxML (SCP) and ASTRAL (SCP) topologies reveal similar relationships (**Figure 3** and **Supplementary Figures 1**, **2**), with a few exceptions of infraspecific accessions. Support values in these two topologies are significantly higher than in the SSP in the mainly Colombian species (**Table 2**). Both samplings are not equivalent, as the SSP includes more species and the SCP includes more and complementary accessions for the poorly supported and less studied clades *parvifolia*, *lindeniana*, *simplex*, and *linearis*; only *A. bio* and *A. cogollo* were left unsampled altogether.

There were several important topologic dissimilarities between the trees obtained from different sequencing strategies, with the most relevant being not the species included in clades but the relationship between clades (i.e., the backbone relationships between the clades, not clade definition). Thus, we chose to discuss the clade relationships obtained by the SCPs, as the resulting supports were higher and sequencing more extensive in the genome, and used the SSP for clade definition (i.e., species placement into clades) because of the inclusion of *A. verrucosa* and *A. grandis*, as well as *A. graminifolia* Galeano & R. Bernal, *A. weberbaueri* Burret, and *A. spicata* Borchs. & R. Bernal in the sampling. An important inconsistency was related to two of the species forming monospecific clades, *A. pilaris* and *A. macroloba*. *A. pilaris* was sampled in both approaches and was recovered as sister to the *parvifolia* complex in the SSP but not in the SCP, where it was recovered as sister to everything but *A. killipii* Burret, the *horrida* and *weberbaueri* clades. In the SCP, *A. macroloba* was placed as sister to both the *lindeniana* and the *linearis* clades, whereas in the SSP it was placed as sister only to the *linearis* clade. Despite these differences, the strategies were complementary by providing species assignment into the clades (SSP) plus relationships between the clades (SCP).

Using the two different calibration schemes for the SSP and SCP resulted in similar divergence age estimates, concentrating in the Miocene-Pliocene. The SNAPP analyses resulted in an estimated divergence time between the *lindeniana* and *linearis* clades of 2-1 million years before the present, and

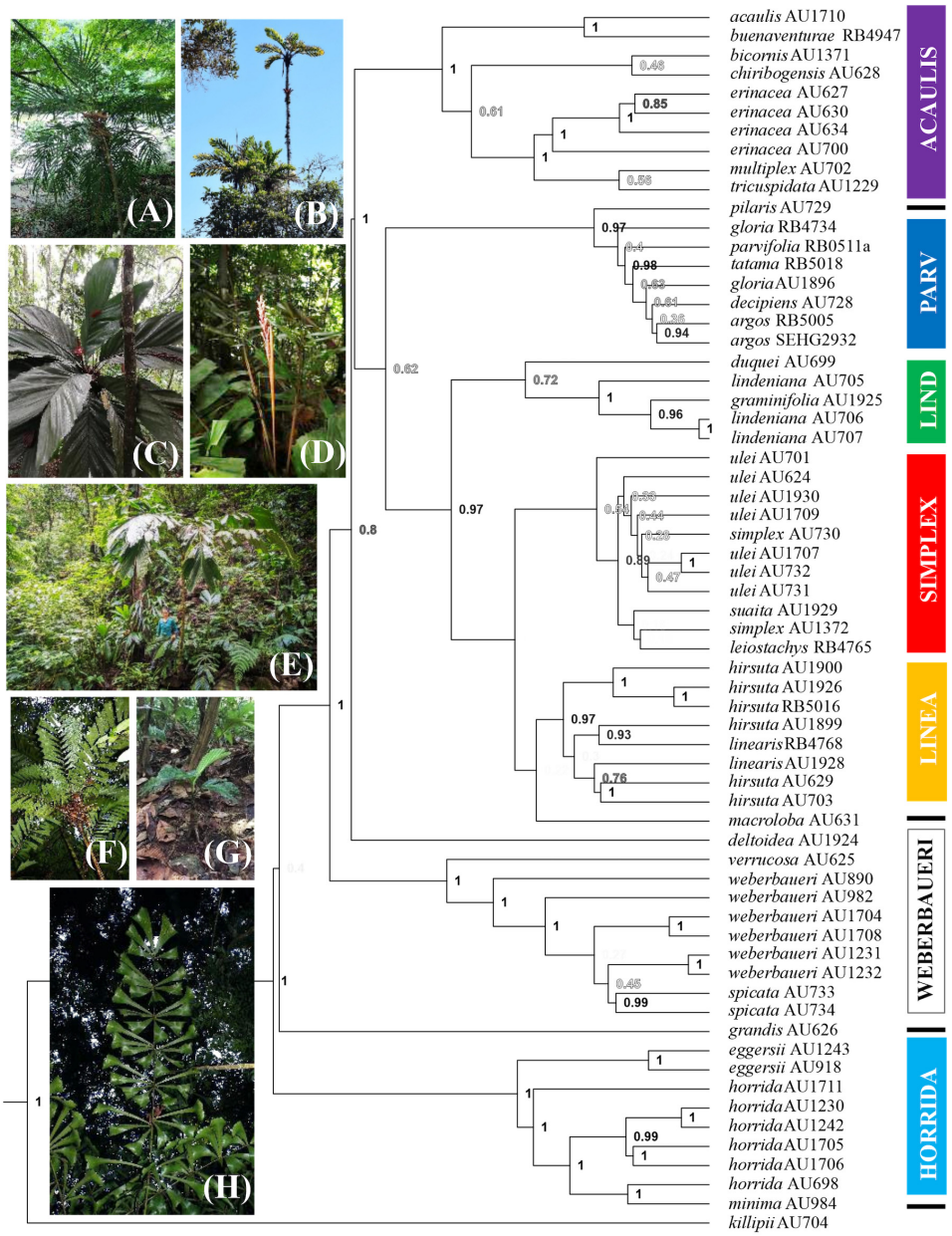

FIGURE 2 | Bayesian phylogeny of the genus *Aiphanes* by Sanger sequencing. Colors of clades marked by boxes follow online **Supplementary Figures 1–4, 6** and **Figure 5**; the lines indicate species not assigned to clades; support values as posterior probabilities are shown in shades of gray with higher support values in black; epithets are followed by voucher code. Photos show **(A)** *Aiphanes argos* (*parvifolia* clade): habit; **(B)** *A. concinna* (*lindeniana* clade): habit; **(C)** *A. cogollo* (*parvifolia* clade): crown; **(D)** *A. bio* (*parvifolia* clade): inflorescence; **(E)** *A. hirsuta* (*linearis* clade): habit; **(F)** *A. leiostachys* (*simplex* clade): crown; **(G)** *A. macroloba*: habit; **(H)** *A. killipii*: funneled pinnae. Photographs: **(C)** by Alvaro Cogollo, **(D)** by Camilo Flórez, **(E)** by Felipe Mesa.

tree inconsistencies regarding the relationship among the four included species (**Supplementary Table 4**).

Geographical Distribution and Overlap Between Same-Clade Species

The first branching *A. killipii* has a narrow distribution in the Eastern Cordillera of Colombia. Further splits in the phylogeny produce the widely distributed *horrida* clade, spanning the Tropical Andes and the Caribbean, a *weberbaueri* clade from mostly Peru and Ecuador, the Colombia and Ecuador Pacific *acaulis* clade, and then the four most nested and predominantly Colombian Andean *linearis*, *lindeniana*, *parvifolia*, and *simplex* clades.

The species of *Aiphanes* present a very small overlap of distribution ranges, with most species in clades showing a

TABLE 2 | Support values (as posterior probabilities) of *Aiphanes* clades for the Bayesian and ASTRAL analyses, and in Bootstrap values for RAxML.

Clades	Clade Support		
	Sanger sequence Bayesian	Sequence Capture ASTRAL	Sequence Capture RAxML
acaulis	1	1	100
parvifolia + *pilaris*	0.97	NA	NA
parvifolia	0.4	1	100
lindeniana	0.72	1	100
deltoidea	NA	NA	NA
horrida	1	1	100
grandis	NA	Not studied	Not studied
linearis	0.97	1	100
killipii	NA	1	100
simplex	0.54	1	100
macroloba	NA	1*	100*
pilaris	NA	NA	NA
weberbaueri	1	Not studied	Not studied

NA, not applicable, meaning not recovered within a clade but as a grade. Asterisk stands for support for a single species sampled with more than one accession.

less than 10% overlap. Average range overlap was highest in the *deltoidea* and *linearis* clades (38 and 31%, respectively), followed by the *acaulis* and *lindeniana* clades (18 and 16%, respectively), and with the other clades showing a less than 10% average overlap. Most of the overlapped geographical ranges include a species that is microendemic (known from one or a few close localities) embedded in a range of a widely distributed species (i.e., 99% of *A. spicata's* range is within *A. weberbaueri's* but only 1% vice versa). Overlap of 31–92% did occur between widely distributed Andean sister species pairs: *A. concinna*/*A. lindeniana*, and *A. hirsuta*/*A. linearis*. The geographical distributions are available in **Supplementary Figure 3** (distribution point maps) and **Supplementary Figure 4** (SDMs used to estimate geographical overlap); the table with all the paired overlap estimates is available in **Supplementary Table 5**.

Evolution of the Climatic Niche in *Aiphanes*

Temperature

Figure 4 and **Supplementary Figure 5** show the mean annual temperature (MAT, or Bio1) reconstruction throughout the Bayesian tree. The genus is and has been adapted to temperatures of ca. 18°C. This MAT is not typical of the lowlands but of mid-elevation forests, between 1,000 and 2,000 m under current climate. Several times independently in the phylogeny, *Aiphanes* species adapt to either colder (close to 11°C of MAT in *A. concinna* H.E. Moore, *A. pilaris*, and *A. ulei* Burret, shown in blue) or warmer MAT (near 26°C, as *A. acaulis, A. horrida*, and *A. deltoidea* Burret, shown in red). The ancestral states of the earliest nodes in *Aiphanes* indicate intermediate values for MAT. Isothermality (Bio3) shows a pattern similar to that of MAT. The two higher clades reconstructed under the SCP (i.e., *lindeniana* + *linearis* + *macroloba* and *acaulis* + *simplex* + *parvifolia*) have lowland, midland, and highland species, covering the full topographic gradient for the whole genus and thus covering the gradient for both MAT and other indicative variables such as isothermality and temperature of the coldest month.

Precipitation

Figure 4 and **Supplementary Figure 5** show that mean annual precipitation (MAP, or Bio12) is in the 2,000–4,000 mm/yr range, with high elevation species adapting to less MAP above 1,700 mm/yr and few coastal or lowland species adapting to higher MAP under 6,000 mm/yr. The ancestral state for all *Aiphanes* for MAP is close to 2,000 mm/yr. Precipitation seasonality (Bio15: the differences in rainfall occurring throughout the year), varies only in a few species in the *acaulis, horrida*, and *parvifolia* clades, which indicates that *Aiphanes* mostly occurred in less seasonal ecosystems in terms of rainfall. Annual precipitation and precipitation seasonality varied inversely, with progressive evolution of the genus toward higher annual precipitation and lower precipitation seasonality. Only two species from the Colombian Choco occur in areas with high annual precipitation of around 6,000 mm/yr: *A. buenaventurae* R. Bernal and Borchs. and *A. acaulis*. Only one species occurs in areas with high precipitation seasonality: *A. eggersii* Burret.

Figure 4 shows the breadth of occupancy by species (rows) and clades (colors from SSP and SCPs) for three climatic variables that are representative of annual variation (mean annual temperature and precipitation, MAT and MAP, and precipitation seasonality). MAT shows segregation between species, whereas MAP and precipitation seasonality tend to be closer between species in clades. Thermal occupation is divided among species of each clade, as can be seen in the "laddered" distribution of the boxes within each color-coded clade.

Climatic Niche Overlap Between Same-Clade Species

The species pair niche comparisons yield significant differences for most same-clade species pairs (**Figure 5** and **Table 3**; refer to **Supplementary Figure 6** for niche volume overlap projections on the two first principal components encompassing 75% of variance and **Supplementary Table 5** for Schoener's *D* index for all species pairs in the clades). Schoener's *D* indices for all species pairs were relatively low, with a minimum of 0, a maximum of 0.65, mean = 0.08, and median = 0.01). The geographically overlapped same-clade species pair (*A. concinna*/*A. lindeniana*) also have significantly overlapped niches from the VIF variable set (Schoener's $D = 0.65$). Climatic niche overlap of species pairs, as measured by Schoener's *D*, is highest in the *acaulis, lindeniana*, and *parvifolia* clades. Niche overlap is, in many cases, related to geographical overlap and is most common between species pairs where at least one species has a very narrow distribution (is only known from one or few localities). Pearson correlation between the average Schoener's *D* value per clade and the age of the clade resulted in a weak negative correlation (-0.349) that is not statistically significant (p-value = 0.4425).

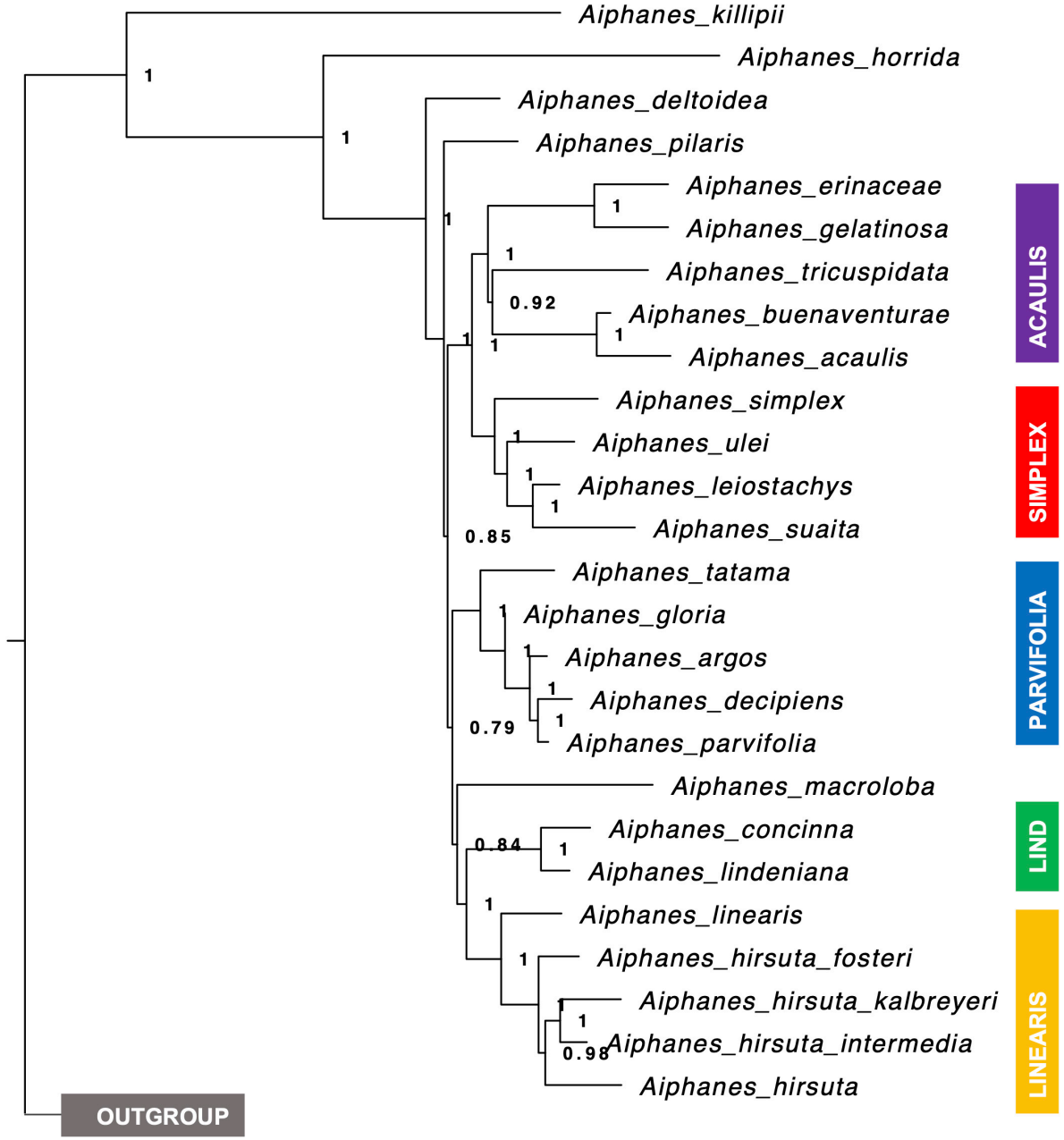

FIGURE 3 | Species tree of *Aiphanes* where tips are represented by a species (or subspecies) and not by each sampled individual based on 515 genes. Node branches show local PP support, and clades are color-coded to match **Figure 2**.

Phylogenetic Signal of Climatic Variables

The phylogenetic signal tests (**Table 4** and **Supplementary Figure 7**) show that the two principal components of temperature- and precipitation-related variables are best explained by different evolutionary models according to the Akaike Information Criterion (AIC), as shown in **Table 4**. The temperature-related variables exhibit less phylogenetic signal than the precipitation-related variables, shown by Blomberg's K (Blomberg et al., 2003) and Pagel's lambda (Pagel, 1999). This is true both for the SSP and the SCP (RAxML). The low values of both indicators (of phylogenetic signal of MAP) show that species of different clades converge to living at similar temperatures.

DISCUSSION

Our complementary taxon and sequence sampling methods allowed us to cover the different depths of the phylogenetic history of *Aiphanes*. The Sanger phylogeny (**Figure 2**) included more species and provided sufficient support for the inclusion

of species into clades, whereas the sequence capture phylogenies (**Figure 3** and **Supplementary Figures 1**, **2**) densely sampled several poorly known species and clades that the Sanger sequence failed to resolve (i.e., the subspecies of *A. hirsuta*, the differentiation between *A. concinna* and *A. lindeniana*). Also, the Sanger sequencing phylogeny did not support the relationships between clades, whereas the sequence capture phylogeny provided a better-resolved backbone for between-clade relationships (**Table 2**). Therefore, we encourage this mixed approach when (1) sampling with more genome-wide methods is not available for all the targeted taxa and (2) different phylogenetic depths are relevant to the study. At the micro-macroevolutionary interphase where populations, subspecies, and species are being sampled, the Sequence Capture approach was most important. We base the following paragraphs of the current section on these complementary results. We use these phylogenetic hypotheses to discuss the roles of geographical and climatic niche evolution in fostering divergence in the genus.

Diversification in *Aiphanes*

The earliest diverging members of the genus (except *A. minima*) occur today in mid-elevation mesic environments (*A. killipii, A. horrida,* and *A. eggersii*). Diversification in the genus appears to have occurred recurrently in different temperatures (**Figure 6A**), and to have been related to the early conquest of areas with higher precipitation and lower precipitation seasonality, and the conservation of this trait (**Figures 4B,C**, **6B,C** and **Supplementary Figure 5**). These precipitation-related adaptations were conserved both at higher elevations in the northern Andes (*linearis* and *lindeniana* clades) and at lower elevations in the *cis-* and *trans-*Andean lowlands (*acaulis* clade in Choco and *A. ulei* and *A. deltoidea* in the Amazon). This is shown by the high phylogenetic signal of the joint precipitation-related variables but not of the temperature-related variables reflected in the lower values of Blomberg's K and Pagel's Lambda for the former variable set (**Table 3** and refer to **Supplementary Figure 7**).

Both the geographic distribution of species and clades and the climatic evolution of the genus highlight its close relationship with Andean mountain building from the origin of the genus, as shown by the character state reconstruction indicating mid-elevation temperatures from the ancestor of the genus and subsequent nodes (**Figure 4**). Also, only *A. hirsuta* reaches Panama and Costa Rica, and *A. minima* Burret the West Indies; the rest of the species are on the Andean Mountains or surrounding lowlands. The mean annual temperature of the nodes in *Aiphanes* is reconstructed as mild, which can be understood as mid-elevations at 1,000–2,000 m. The MAT during the late Miocene to Pliocene, when *Aiphanes* diversified (Eiserhardt et al., 2011b; this study), was probably warmer than today (Zachos et al., 2001; Hansen et al., 2013); thus we expect mild temperatures to be related to paleoelevation. Furthermore, as expected by increasing mountain building, high elevation species belong to the more nested clades of *lindeniana* and *linearis*, as well as the species *A. spicata* and *A. pilaris*.

The progression from mid to higher elevations probably follows the process of mountain building and points at mid-elevation areas in the Colombian Andes (possibly the Eastern Cordillera where *A. killipii* is found today and is near two other microendemic species from different clades, *A. graminifolia* and *A. suaita*) already available during the Miocene (Anderson et al., 2015). Despite the absence of paleoelevation constraints for the Central and Western Cordilleras, thermochronological and geomorphological considerations suggest that these Cordilleras

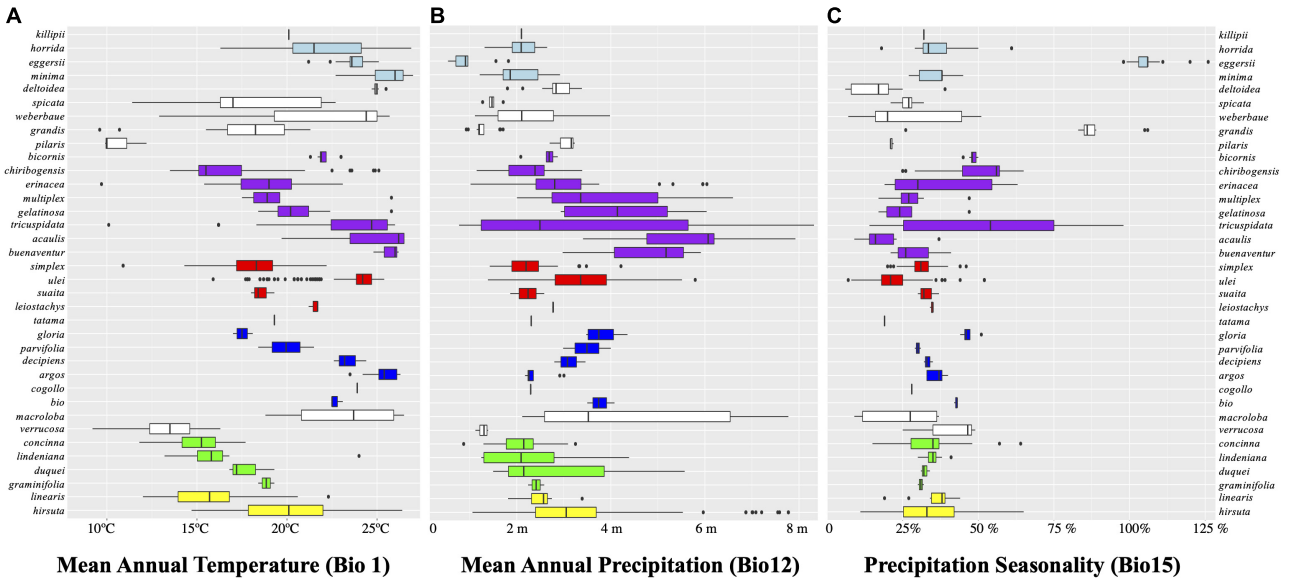

FIGURE 4 | Boxplots for three of the variables evaluated grouped by clades indicated by the same colors from **Figure 2**. **(A)** Bio1, mean annual temperature in Celsius degrees; **(B)** Bio12, mean annual precipitation in 1 year in kg m^{-2}; **(C)** Bio15, precipitation seasonality in meters of rain in kg m^{-2} (i.e., the standard deviation of monthly precipitation estimates is expressed as a percentage of their mean).

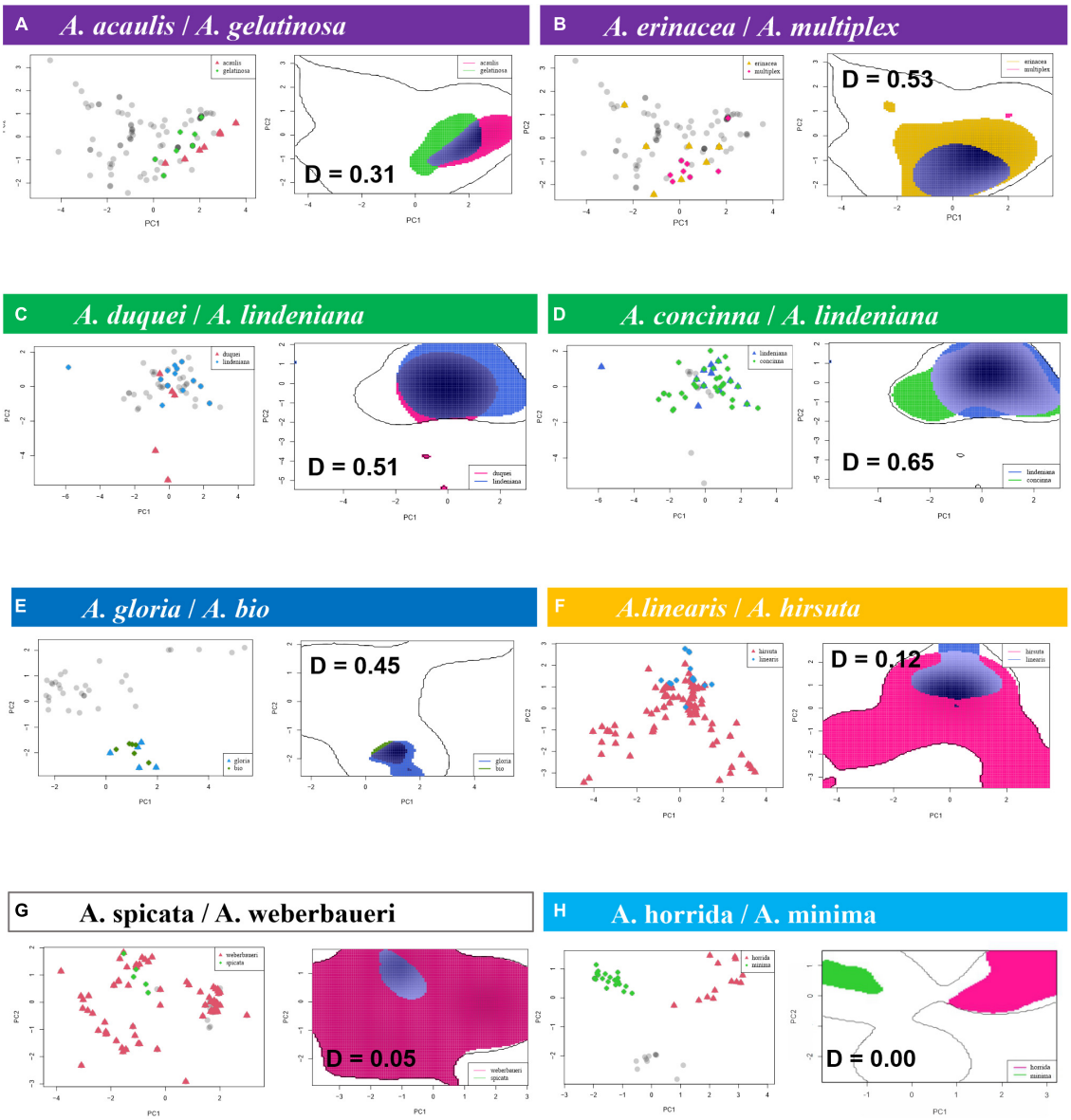

FIGURE 5 | Niche overlap, projected on the two first components of the variance inflation factor variable principle component analyses, can occur in relation to a wide variety of geographical species ranges: **(A,E)** between two microendemics; **(C,G)** between a microendemic and a widely distributed species, **(B,D,F)** between two widely distributed species occurring in the **(B)** lowlands, **(D)** highlands, or **(F)** throughout the elevation gradient; **(H)** two widely distributed same-clade species with entirely non-overlapping niches. Schoener's *D* in bold, including the highest level of observed overlap between same-clade species (*D* = 0.65 in the *lindeniana* clade) and the lowest possible value (*D* = 0) observed for many same species pairs and all species pairs in the *horrida* clade. The same-clade species here are understood as belonging to the same outline clade (in colors).

may have also experienced exhumation and uplift since the Miocene (León et al., 2018; Noriega-Londoño et al., 2020). Nevertheless, the concentrated origin of several early-diverging species in the Eastern Cordillera suggests that this area was either uplifted or connected earlier in time. The high-elevation *A. linearis* from the Western and Eastern Cordilleras is nested in the phylogeny and is among the last to form. Diversification in the Colombian cordilleras and the Choco (the *acaulis, parvifolia, lindeniana, linearis,* and *simplex* clades), as well as tendency to gradually occupy areas with higher precipitation and less seasonality and a wide variety of temperatures, hints at how the evolution of *Aiphanes* and the North Andean chains are intertwined.

In simple terms, the tracking of these palms of continuously moist environments allowed them to diversify, occupying different temperatures available at the places where they became geographically isolated by mountain building (**Figure 7**; the elevation ranges for all species are available

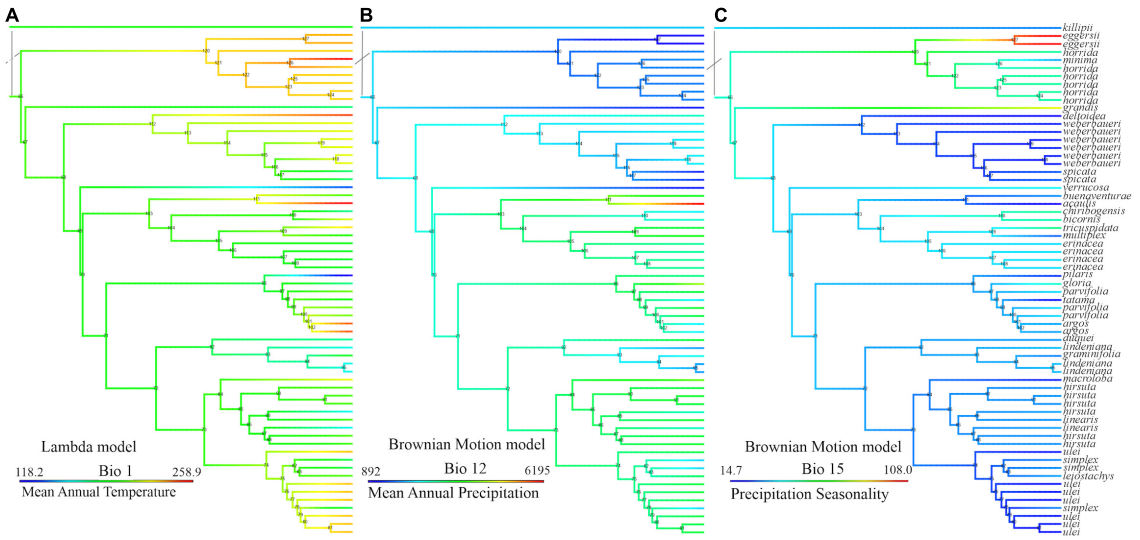

FIGURE 6 | Climatic variable reconstruction according to the best-fitting model of continuous trait evolution by the Akaike Information Criterion, values reconstructed from all curated species occurrences, and CHELSA climatic layers on the Sanger sequence phylogeny (SSP); **(A)** mean annual temperature by the Lambda model, values in deciCelsius, **(B)** mean annual precipitation by Brownian Motion, values in kg $m^{-2}/10$, and **(C)** precipitation seasonality by Brownian motion, values in percentage of monthly variation over annual mean.

TABLE 3 | Niche and geographical range overlap of species pairs in clades; for complete list of species pairs, refer to **Supplementary Table 5**; here, we only show species pairs where Schoener's *D* index or *p*-values for equivalency and similarity indicate that climatic niches are not different; we also show their corresponding geographical range overlaps (ROs) in km^2.

Clade	Species pairs	Schoener's *D* Index	*p*-value *equivalency*	*p*-value *similarity*	Area sp. 1 (km^2)	Area sp. 2 (km^2)	RO sp. 1 vs. 2	RO sp. 2 vs. 1
ACAULIS	*acaulis vs. gelatinosa*	0.3084	0.91089	0.17822	12505.7	15307.4	0.07	0.09
ACAULIS	*buenaventurae vs. gelatinosa*	0.2305	0.9109	0.0396	913.7	15307.4	0.00	0.00
ACAULIS	*multiplex vs. gelatinosa*	0.2104	0.9802	0.2376	27022.8	15307.4	0.43	0.76
ACAULIS	*erinacea vs. gelatinosa*	0.2605	0.8317	0.1485	64697.2	15307.4	0.00	0.84
ACAULIS	*erinacea vs. multiplex*	0.5246	0.5149	0.0099	64697.2	27022.8	0.34	0.81
LINDENIANA	*duquei vs. lindeniana*	0.5093	0.4554	0.3366	608.6	148344.8	0.36	0.00
LINDENIANA	*duquei vs. concinna*	0.4422	0.2475	0.2178	608.6	79175.9	0.24	0.00
LINDENIANA	*lindeniana vs. concinna*	0.6455	0.1287	0.0693	148344.8	79175.9	0.49	0.93
PARVIFOLIA	*cogollo vs. argos*	0.3005	0.7129	0.0594	294.8	1652.8	0.53	0.09
PARVIFOLIA	*gloria vs. bio*	0.4460	0.4753	0.0099	510.0	342.5	0.00	0.00

TABLE 4 | Phylogenetic signal of temperature- and precipitation-related variables on the sequence capture phylogeny (SCP) (RAxML) and Sanger sequence phylogeny (SSP).

Variable used	Representative bio-variable	Model of best fit by AIC and AICc (SSP)	Model of best fit by AIC and AICc (SCP)	K-Statistic of phylogenetic signal (SSP)	K-Statistic of phylogenetic signal (SCP)	Pagel's lambda (SSP)	Pagel's lambda (SCP)
PC1_temp	Bio1	Lambda	Lambda	0.41	0.68	0.46	0.59
PC2_temp	Bio3	Ornstein-Uhlenbeck	Ornstein-Uhlenbeck	0.43	0.68	0	0.59
PC1_prec	Bio15	Brownian Motion	Early Burst	0.71	1.35	1	1
PC2_prec	Bio12	Brownian Motion	Brownian Motion	0.96	1.05	1	1

in **Supplementary Figure 8**). High level of precipitation and low precipitation seasonality have been the backdrop for species persistence in the dynamic Neogene evolution of Northwestern South America (Hoorn et al., 2010), and are conserved traits in the phylogeny. On the other hand, the repeated temperature differentiation in the different clades is the result of the evolving

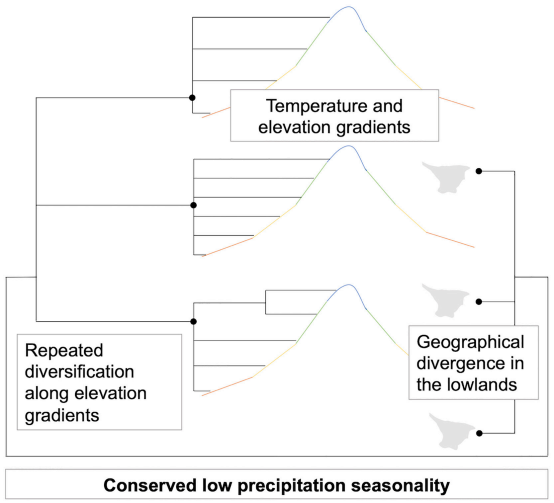

FIGURE 7 | Schematic representation of how *Aiphanes* diversified.

topography, isolating species in valleys and mountains at different elevations.

Geographical divergence is important in different Andean birds (Hazzi et al., 2018) and plants (Vargas, 2018; Vargas et al., 2020) and is reflected in the low mean geographical overlap between species and their relatively small ranges. Also, it can, in some cases, be concomitant with niche adaptation (Benham and Witt, 2016). Indeed, the temperature differentiation here follows a spreading pattern shown by the parallel laddered occupation of different temperatures of same-clade species (**Figure 4**, refer to Bio1 occupation for clades *horrida* in light blue, *acaulis* in purple, *parvifolia* in blue, and *lindeniana* in green). Climatic space differentiation was also shown for an Andean endemic palm genus, *Ceroxylon* (Sanín et al., 2016), and Andean vertebrates (Patterson et al., 1998; Cadena et al., 2011). In the case of *Aiphanes*, topography plays a dual role by isolating populations and then truncating their potential occupation along the temperature gradient (Benham and Witt, 2016).

It is widely acknowledged that realized distributions underestimate species' physiological tolerance (Soberon and Peterson, 2005; Feeley and Silman, 2010), and that physical barriers to dispersal might amplify this effect. The resulting physically constrained extension of species' natural ranges enforces niche truncation (Bush et al., 2018). In *Aiphanes,* the rarely overlapping occupied temperature niches between same-clade species might be the effect of dispersal limitations affecting climatic niche realization because of, in particular, topographic barriers. Thus, our estimation of potential niche occupation, which relies on occurrences, might be ultimately biased by how well the species has succeeded in "sampling" the topographically fragmented landscape. Conversely, it is clear that the occupation of precipitation-related climatic space is much more phylogenetically constrained, as it is conserved despite geographical circumscription.

The overall pattern appears as diversification by geographical isolation in moist areas with different temperatures, which has made *Aiphanes* species-rich on the regional scale but locally rare (Borchsenius and Bernal, 1996) and locally species-poor. Ultimately, this recoils to the genus' aforementioned paucity. Either the relative scarcity of pollinators or fluctuations in their specificity or other biotic factors, such as poor ability to recolonize areas due to poor seedling recruitment (Svenning, 1998), may have limited the population expansion of most species of *Aiphanes*. The reproductive biology of this genus has only been studied in *A. erinacea* and in *A. chiribogensis* Borchs. & Balslev, where hoverflies and gnats/midges, respectively, were reported as the main pollinators in Ecuador (Borchsenius, 1993). Our field observation is that infructescences and ripe fruits are seldom found in wild populations, but this has not been systematically assessed.

This study could be expanded in different ways. First, we lack natural history knowledge of the physiological tolerance of compound variables like energy-water balance in many tropical plant groups. Second, we lack knowledge of actual dispersal abilities mediated by biotic interactions in tropical taxa. Third, we should further understand geographical isolation: is it slope, topographic complexity, or the temporal and spatial extent of a physical barrier that is related to biodiversity patterns determined by isolation? A limitation of this study stems from these knowledge gaps.

Concluding Remarks

As seen in the previous sections, *Aiphanes* provides an example of a diversification pattern by the occupation of the elevation gradient following high precipitation with low seasonality. Isolation into different corners of Andean relief leads to temperature differentiation the same way repeatedly across the different clades, and to high abundance of narrowly distributed species. The phylogenetic conservation of precipitation-related adaptations is coupled with possibly fundamentally larger tolerance to different temperatures. However, wider tolerance to different temperatures is not reflected in species occupation of climatic space (i.e., their realized climatic niches), because it is truncated by physical occupation of mountains and valleys from where the species are unable to disperse. This pattern might be common in Andean plant groups but understudied because of lack of taxonomic understanding, collections, sequencing, and validated distribution data for plant groups that include a high proportion of rare and narrowly distributed species.

In *Aiphanes*, future research should aim at discovering how this often-unnoticeable genus maintains populations and effective reproduction in time despite its elusive presence in mountain forests. This, combined with the expected discovery of additional microendemic species, will help to understand the evolution of this fascinating Andean gem.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/, PRJNA689999.

AUTHOR CONTRIBUTIONS

MS, FB, and RB designed the study. MS, FB, RB, SH-G, HM, and AC collected the data. MS, AG, FB, SC-M, YO, NA, and MP analyzed the data. MS wrote the manuscript. All authors revised, contributed, and approved the manuscript.

FUNDING

This study was funded by the Colciencias grants 173-2016 and 80740-606-2019 to MS and AC.

ACKNOWLEDGMENTS

We thank Juan Sebastián Jaramillo, Julissa Roncal, Vanessa Correa, Lina Bolívar, Dino Tuberquia, Ana Ospina, Ángela Cano, and David Esteban Hernández for their help in the field or for providing samples, Fabián Mejía for extracting the DNA of many of the samples, lab technicians Diana Carmona and Sergio Alzate for their support in keeping the lab running at Universidad CES, and the gardeners at Jardín Botánico del Quindío, where some of the plants sampled in this study are preserved. We especially thank the director of the Biological Collections at Universidad CES (CBUCES), Juliana Cardona, for her energetic and positive support throughout. Some of the samples were collected under the institutional permit resolution No. 0790 granted to Universidad CES by ANLA and others were taken from herbarium specimens.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.881879/full#supplementary-material

REFERENCES

Altamiranda-Saavedra, M., Arboleda, S., Parra, J. L., Townsend Peterson, A., and Correa, M. M. (2017). Potential distribution of mosquito vector species in a primary malaria endemic region of Colombia". *PLoS One* 12:e0179093. doi: 10.1371/journal.pone.0179093

Anderson, V. J., Saylor, J. E., Shanahan, T. M., and Horton, B. K. (2015). Paleoelevation records from lipid biomarkers: application to the tropical Andes. *Bull. Geol. Soc. Am.* 127, 1604–1616. doi: 10.1130/B31105.1

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., Kissling, W. D., Flantua, S. G. A., Bermúdez, M. A., Mulch, A., Muellner-Riehl, A. N., et al. (2018). Geological and climatic influences on mountain biodiversity. *Nat. Geosci.* 11, 718–725. doi: 10.1038/s41561-018-0236-z

Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S. P., Peterson, A. T., et al. (2011). The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecol. Modell.* 222, 1810–1819.

Benham, P. M., and Witt, C. C. (2016). The dual role of Andean topography in primary divergence: functional and neutral variation among populations of the hummingbird, metalluratyrianthina. *BMC Evol. Biol.* 16:22. doi: 10.1186/s12862-016-0595-2

Bernal, R., and Borchsenius, F. (2010). Taxonomic novelties in *Aiphanes* (Palmae) from Colombia and Venezuela. *Caldasia* 32, 117–127.

Bernal, R., Borchsenius, F., Hoyos-Gómez, S. E., Manrique, H. F., and Sanín, M. J. (2019a). A revision of the *Aiphanes parvifolia* complex (Arecaceae). *Phytotaxa* 411, 275–292. doi: 10.11646/phytotaxa.411.4.3

Bernal, R., Castaño, F., and Sanín, M. J. (2019b). A new, overlooked species of *Aiphanes* (Arecaceae) from santander, Colombia. *Phytotaxa* 405, 101–105. doi: 10.11646/phytotaxa.405.2.5

Bernal, R., Hoyos-Gómez, S. E., and Borchsenius, F. (2017). A new, critically endangered species of *Aiphanes* (Arecaceae) from Colombia. *Phytotaxa* 298, 65–70. doi: 10.11646/phytotaxa.298.1.6

Blomberg, S. P., Garland, T., and Ives, A. R. (2003). Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution* 57, 717–745. doi: 10.1111/j.0014-3820.2003.tb00285.x

Bonetti, M. F., and Wiens, J. J. (2014). Evolution of climatic niche specialization: a phylogenetic analysis in amphibians. *Proc. Biol. Sci.* 281:20133229. doi: 10.1098/rspb.2013.3229

Borchsenius, F. (1993). Flowering biology and insect visitation of three Ecuadorean *Aiphanes* species. *Principes* 37, 139–150.

Borchsenius, F., and Bernal, R. (1996). *Aiphanes* (Palmae). *Flora Neotrop.* 70, 1–94.

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

Broennimann, O., Fitzpatrick, M. C., Pearman, P. B., Petitpierre, B., Pellissier, L., Yoccoz, N. G., et al. (2012). Measuring ecological niche overlap from occurrence and spatial environmental data. *Glob. Ecol. Biogeogr.* 21, 481–497. doi: 10.1111/j.1466-8238.2011.00698.x

Bruggeman, J., Heringa, J., and Brandt, B. W. (2009). PhyloPars: estimation of missing parameter values using phylogeny. *Nucleic Acids Res.* 37, W179–W184. doi: 10.1093/nar/gkp370

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

Burret, M. (1932). Die palmengattungen martineziaund *Aiphanes*. Notizblatt des königl. *Botanischen Gartens Und Museums Zu Berlin* 11, 557–577. doi: 10.2307/3995129

Bush, A., Catullo, R. A., Mokany, K., Thornhill, A. H., Miller, J. T., and Ferrier, S. (2018). Truncation of thermal tolerance niches among Australian plants. *Glob. Ecol. Biogeogr.* 27, 22–31. doi: 10.1111/geb.12637

Cadena, C. D., Kozak, K. H., Gómez, J. P., Parra, J. L., Mccain, C. M., Bowie, R. C. K., et al. (2011). Latitude, elevational climatic zonation and speciation in new World vertebrates. *Proc. R. Soc. B Biol. Sci.* 279, 194–201. doi: 10.1098/rspb.2011.0720

Calixto-Pérez, E., Alarcón-Guerrero, J., Ramos-Fernández, G., Dias, P. A. D., Rangel-Negrín, A., Améndola-Pimenta, M., et al. (2018). Integrating expert knowledge and ecological niche models to estimate Mexican primates' distribution. *Primates* 59, 451–467. doi: 10.1007/s10329-018-0673-8

Chamberlain, S., and Boettiger, C. (2017). R Python, and Ruby clients for GBIF species occurrence data. *PeerJ Prepr.* 5:e3304v1. doi: 10.7287/peerj.preprints.3304v1

Chamberlain, S., Barve, V., Mcglinn, D., Oldoni, D., Desmet, P., Geffert, L., et al. (2022). *rgbif: Interface to the Global Biodiversity Information Facility API. R Package Version 3.7.2*.

Chazot, N., De-Silva, D. L., Willmott, K. R., Freitas, A. V. L., Lamas, G., Mallet, J., et al. (2018). Contrasting patterns of Andean diversification among three diverse clades of Neotropical clearwing butterflies. *Ecol. Evol.* 8, 3965–3982. doi: 10.1002/ece3.3622

Chazot, N., Willmott, K. R., Condamine, F. L., De-Silva, D. L., Freitas, A. V. L., Lamas, G., et al. (2016). Into the Andes: multiple independent colonizations drive montane diversity in the neotropical clearwing butterflies godyridina. *Mol. Ecol.* 25, 5765–5784. doi: 10.1111/mec.13773

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

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

de 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

Di Cola, V., Broennimann, O., Petitpierre, B., Breiner, F. T., D'Amen, M., Randin, C., et al. (2017). Ecospat: an R package to support spatial analyses and modeling of species niches and distributions. *Ecography* 40, 774–787. doi: 10.1111/ecog.02671

Dransfield, J., Uhl, N. W., Asmussen, C. B., Baker, W. J., Harley, M. M., and Lewis, C. E. (2008). *Genera Palmarum-The Evolution and Classification of the Palms*. Richmond: Kew Publishing, 732.

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

Eiserhardt, W. L., Pintaud, J.-C., Asmussen-Lange, C., Hahn, W. J., Bernal, R., Balslev, H., et al. (2011a). Phylogeny and divergence times of bactridinae (arecaceae, palmae) based on plastid and suclear DNA Sequences. *Taxon* 60, 485–498. doi: 10.1002/tax.602016

Eiserhardt, W. L., Svenning, J. C., Kissling, W. D., and Balslev, H. (2011b). Geographical ecology of the palms (arecaceae): determinants of diversity and distributions across spatial scales. *Ann. Bot.* 108, 1391–1416. doi: 10.1093/aob/mcr146

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., and Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Divers. Distrib.* 17, 43–57. doi: 10.1111/j.1472-4642.2010.00725.x

Escobar, D., Jojoa, L. M., Díaz Sánchez, S. R., Rudas, E., Albarracín, R., Gómez, J., et al. (2016). *Georreferenciación de Localidades: Una Guía de Referencia Para Colecciones Biológicas.Instituto de Investigación de Recursos Biológicos Alexander von Humboldt –Instituto de Ciencias Naturales*. Bogotá: Universidad Nacional de Colombia, 144.

Feeley, K. J., and Silman, M. R. (2010). Biotic attrition from tropical forests correcting for truncated temperature niches. *Glob. Change Biol.* 16, 1830–1836. doi: 10.1111/j.1365-2486.2009.02085.x

Flantua, S. G. A., O'Dea, A., Onstein, R. E., Giraldo, C., and Hooghiemstra, H. (2019). The flickering connectivity system of the north Andean páramos. *J. Biogeogr.* 46, 1808–1825. doi: 10.1111/jbi.13607

Galeano, G., and Bernal, R. (2010). *Palmas de Colombia: Guía de Campo*. Bogotá: Universidad Nacional de Colombia.

Galeano, G., Bernal, R., and Figueroa Cardozo, Y. (2015). *Plan de conservación, manejo y uso sostenible de las palmas de Colombia*. Bogotá: Universidad Nacional de Colombia-Ministerio de Ambiente y Desarrollo Sostenible.

Gaut, B. S., Morton, B. R., Mccaig, B. C., and Clegg, M. T. (1996). Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene RbcL. *Proc. Natl. Acad. Sci. U.S.A.* 93, 10274–10279. doi: 10.1073/pnas.93.19.10274

Graham, C. H., Carnaval, A. C., Cadena, C. D., Zamudio, K. R., Roberts, T. E., Parra, J. L., et al. (2014). The origin and maintenance of montane diversity: integrating evolutionary and ecological processes. *Ecography* 37, 711–719. doi: 10.1111/ecog.00578

Graham, C., Ron, S. R., Santos, J. C., Schneider, C. J., and Moritz, C. (2004). Integrating phylogenetics and environmental niche models to explore speciation mechanisms in dendrobatid frogs. *Evolution* 58, 1781–1793. doi: 10.1111/j.0014-3820.2004.tb00461.x

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

Hansen, J., Sato, M., Russell, G., and Kharecha, P. (2013). Climate sensitivity, sea level and atmospheric carbon dioxide. *Philos. Trans. R. Soc.* 371:20120294. doi: 10.1098/rsta.2012.0294

Harvey, P. H., and Pagel, M. (1991). *The Comparative Method in Evolutionary Biology*. Oxford: Oxford University Press.

Hazzi, N. A., Moreno, J. S., Ortiz-Movliav, C., and Palacio, R. D. (2018). Biogeographic regions and events of isolation and diversification of the endemic biota of the tropical Andes. *Proc. Natl. Acad. Sci. U.S.A.* 115, 7985–7990. doi: 10.1073/pnas.1803908115

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

Ho, L. S. T., and Ané, C. (2014). A linear-time algorithm for gaussian and non-gaussian trait evolution models. *Syst. Biol.* 63, 397–408. doi: 10.1093/sysbio/syu005

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* 330, 927–931. doi: 10.1126/science.1194585

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., et al. (2017). Climatologies at high resolution for the earth's land surface areas. *Sci. Data* 4:170122. doi: 10.1038/sdata.2017.122

Kristiansen, T., Svenning, J. C., Pedersen, D., Eiserhardt, W., Grández, C., and Balslev, H. (2011). Local and regional palm (Arecaceae) species richness patterns and their cross-scale determinants in the western Amazon. *J. Ecol.* 99, 1001–1015. doi: 10.1111/j.1365-2745.2011.01834.x

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

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

León, S., Cardona, A., Parra, M., Sobel, E. R., Jaramillo, J. S., Glodny, J., et al. (2018). Transition from collisional to subduction-related regimes: an example from neogene Panama-Nazca-South America interactions. *Tectonics* 37, 119–139. doi: 10.1002/2017TC004785

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

Liu, L., Xi, Z., Wu, S., Davis, C. C., and Edwards, S. V. (2015b). 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). Estimating species phylogenies using coalescence times among sequences. *Syst. Biol.* 58, 468–477. doi: 10.1093/sysbio/syp031

Loiseau, O., Olivares, I., Paris, M., de La Harpe, M., Weigand, A., Koubínová, D., et al. (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

Luebert, F., and Weigend, M. (2014). Phylogenetic insights into Andean plant diversification. *Front. Ecol. Evol.* 2:27. doi: 10.3389/fevo.2014.00027

Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., et al. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* 47, W636–W641. doi: 10.1093/nar/gkz268

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

Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb.* 2010:db.rot5448. doi: 10.1101/pdb.prot5448

Miller, M. A., Pfeiffer, W., and Schwartz, T. (2010). "Creating the CIPRES science gateway for inference of large phylogenetic trees," in *Proceedings of the 2010 Gateway Computing Environments Workshop (GCE)*, Piscataway, NJ: IEEE.

Mirarab, S., and Warnow, T. (2015). ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* 31, i44–i52. doi: 10.1093/bioinformatics/btv234

Mirarab, S., 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

Mittermeier, R. A., Robles Gil, P., Hoffmann, M., Pilgrim, J., Brooks, T., Mittermeier, C. G., et al. (2004). *Hotspots Revisited: Earth's Biologically Richest and Most Endangered Terrestrial Ecoregions*. Mexico: CEMEX.

Mittermeier, R. A., Turner, W. R., Larsen, F. W., Brooks, T. M., and Gascon, C. (2011). "Global biodiversity conservation: the critical role of hotspots," in *Biodiversity Hotspots*, eds F. E. Zachos and J. C. Habel (Heidelberg: Springer Publishers), 3–22.

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

Muellner-Riehl, A. N., Schnitzler, J., Kissling, W. D., Mosbrugger, V., Rijsdijk, K. F., Seijmonsbergen, A. C., et al. (2019). Origins of global mountain plant biodiversity: testing the 'mountain-geobiodiversity hypothesis'. *J. Biogeogr.* 46, 2826–2838. doi: 10.1111/jbi.13715

Myers, N., Mittermeier, R., Mittermeier, C., Da Fonseca, G., and Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature* 403, 853–858. doi: 10.1038/35002501

Naimi, B., Hamm, N. A. S., Groen, T. A., Skidmore, A. K., and Toxopeus, A.-G. (2014). Where is positional uncertainty a problem for species distribution modelling? *Ecography* 37, 191–203. doi: 10.1111/j.1600-0587.2013.00205.x

Neves, D. M., Dexter, K. G., Baker, T. R., Coelho de Souza, F., Oliveira-Filho, A. T., Queiroz, L. P., et al. (2020). Evolutionary diversity in tropical tree communities peaks at intermediate precipitation. *Sci. Rep.* 10:1188. doi: 10.1038/s41598-019-55621-w

Noriega-Londoño, S., Restrepo-Moreno, S. A., Vinasco, C., Bermúdez, M. A., and Min, K. (2020). Thermochronologic and geomorphometric constraints on the cenozoic landscape evolution of the northern Andes: northwestern central cordillera, Colombia. *Geomorphology* 351:106890. doi: 10.1016/j.geomorph.2019.106890

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 earth: a new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *BioScience* 51, 933–938. doi: 10.1641/0006-35682001051[0933:TEOTWA]2.0.CO;2

Orme, C. D. L., Freckleton, R. P., Thomas, G. H., Petzoldt, T., and Fritz, S. A. (2012). *Caper: Comparative Analyses of Phylogenetics and Evolution in R. R Package Version 0.5.2*. Austria: R Foundation for Statistical Computing.

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature* 401, 877–884. doi: 10.1038/44766

Paradis, E. (2013). Molecular dating of phylogenies by likelihood methods: a comparison of models and a new information criterion. *Mol. Phylogenet. Evol.* 67, 436–444. doi: 10.1016/j.ympev.2013.028

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

Patterson, B. D., Stotz, D. F., Solari, S., Fitzpatrick, J. W., and Pacheco, V. (1998). Contrasting patterns of elevational zonation for birds and mammals in the Andes of southeastern peru". *J. Biogeogr.* 25, 593–607. doi: 10.1046/j.1365-2699.1998.2530593.x

Phillips, S. J., Dudík, M., and Schapire, R. E. (2017). *Maxent Software for Modeling Species Niches and Distributions (Version 3.4.1)*. Available online at: https://biodiversityinformatics.amnh.org/open_source/maxent/ (accessed August, 2021).

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Rabiee, M., Sayyari, E., and Mirarab, S. (2019). Multi-allele species reconstruction using ASTRAL. *Mol. Phylogenet. Evol.* 130, 286–296. doi: 10.1016/j.ympev.2018.10.033

Rahbek, C., Borregaard, M. K., Colwell, R. K., Dalsgaard, B., Holt, B. G., Morueta-Holme, N., et al. (2019). Humboldt's enigma: what causes global patterns of mountain biodiversity? *Science* 365, 1108–1113. doi: 10.1126/science.aax0149

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

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Salm, R., Salles, N. V., Alonso, W. J., and Schuck-Paim, C. (2007). Cross-scale determinants of palm species distribution. *Acta Amaz.* 37, 17–25. doi: 10.1590/S0044-59672007000100002

Sanín, M. J., Kissling, W. D., Bacon, C., 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.* 16, 303–317. doi: 10.1111/boj.12419

Schnitzler, J., Graham, C. H., Dormann, C. F., Schiffers, K., and Linder, H. P. (2012). Climatic niche evolution and species diversification in the cape flora, South Africa. *J. Biogeogr.* 39, 2201–2211.

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

Soberon, J., and Peterson, A. T. (2005). Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodivers. Inform.* 2, 1–10. doi: 10.17161/bi.v2i0.4

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

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

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

Svenning, J. C. (1998). The effect of land-use on the local distribution of palm species in an Andean rain forest fragment in northwestern Ecuador. *Biodivers. Conserv.* 7, 1529–1537. doi: 10.1023/A:1008831600795

Svenning, J. C., Borchsenius, F., Bjorholm, S., and Balslev, H. (2008). High tropical net diversification drives the new world latitudinal gradient in palm (arecaceae) species richness. *J. Biogeogr.* 35, 394–406.

Svenning, J. C., Harlev, D., Sorensen, M., and Balslev, H. (2009). Topographic and spatial controls of palm species distributions in a montane rain forest, southern Ecuador. *Biodivers. Conserv.* 18, 219–228.

Vargas, O. M. (2018). Reinstatement of the genus piofontia: a PHYLOGENOMIC-based study reveals the biphyletic nature of diplostephium (asteraceae: astereae). *Syst. Bot.* 43, 485–496. doi: 10.1600/036364418X697210

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

Warren, D. L., Glor, R. E., and Turelli, M. (2008). Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. *Evolution* 62, 2868–2883. doi: 10.1111/j.1558-5646.2008.00482.x

Zachos, J., Pagani, H., Sloan, L., Thomas, E., and Billups, K. (2001). Trends, rhythms, and aberrations in global climate 65 ma to present. *Science* 292, 686–693. doi: 10.1126/science.1059412

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

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

Copyright © 2022 Sanín, Borchsenius, Paris, Carvalho-Madrigal, Gómez Hoyos, Cardona, Arcila Marín, Ospina, Hoyos-Gómez, Manrique and Bernal. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 04 July 2022
doi: 10.3389/fpls.2022.883151

Revised Species Delimitation in the Giant Water Lily Genus *Victoria* (Nymphaeaceae) Confirms a New Species and Has Implications for Its Conservation

Lucy T. Smith[1†], Carlos Magdalena[1†], Natalia A. S. Przelomska[1,2†], Oscar A. Pérez-Escobar[1*], Darío G. Melgar-Gómez[3], Stephan Beck[4], Raquel Negrão[1], Sahr Mian[1], Ilia J. Leitch[1], Steven Dodsworth[5], Olivier Maurin[1], Gaston Ribero-Guardia[6], César D. Salazar[7], Gloria Gutierrez-Sibauty[3], Alexandre Antonelli[1,8,9] and Alexandre K. Monro[1*]*

[1] Royal Botanic Gardens, Kew, Richmond, United Kingdom, [2] National Museum of Natural History, Smithsonian Institution, Washington, DC, United States, [3] Herbario German Coimbra Sanz, Jardín Botánico Municipal de Santa Cruz de la Sierra, Santa Cruz de la Sierra, Bolivia, [4] Herbario Nacional de Bolivia, Universidad Mayor de San Andrés, La Paz, Bolivia, [5] School of Biological Sciences, University of Portsmouth, Portsmouth, United Kingdom, [6] La Rinconada Ecoparque, Santa Cruz, Urbari, Bolivia, [7] Calle 11 Norte #24, Urbari, Bolivia, [8] Gothenburg Global Biodiversity Centre, Department of Biological and Environmental Sciences, University of Gothenburg, Gothenburg, Sweden, [9] Department of Plant Sciences, University of Oxford, Oxford, United Kingdom

Reliably documenting plant diversity is necessary to protect and sustainably benefit from it. At the heart of this documentation lie species concepts and the practical methods used to delimit taxa. Here, we apply a total-evidence, iterative methodology to delimit and document species in the South American genus *Victoria* (Nymphaeaceae). The systematics of *Victoria* has thus far been poorly characterized due to difficulty in attributing species identities to biological collections. This research gap stems from an absence of type material and biological collections, also the confused diagnosis of *V. cruziana*. With the goal of improving systematic knowledge of the genus, we compiled information from historical records, horticulture and geography and assembled a morphological dataset using citizen science and specimens from herbaria and living collections. Finally, we generated genomic data from a subset of these specimens. Morphological and geographical observations suggest four putative species, three of which are supported by nuclear population genomic and plastid phylogenomic inferences. We propose these three confirmed entities as robust species, where two correspond to the currently recognized *V. amazonica* and *V. cruziana*, the third being new to science, which we describe, diagnose and name here as *V. boliviana* Magdalena and L. T. Sm. Importantly, we identify new morphological and molecular characters which serve to distinguish the species and underpin their delimitations. Our study demonstrates how combining different types of character data into a heuristic, total-evidence approach can enhance the reliability with which biological diversity of morphologically challenging groups can be identified, documented and further studied.

Keywords: *Victoria*, heuristic species concept, morphology, population genomics, Victorian era, Mamoré River, molecular diagnosis of species, divergence times

OPEN ACCESS

Edited by:
Gerald Matthias Schneeweiss, University of Vienna, Austria

Reviewed by:
Jinming Chen, Wuhan Botanical Garden (CAS), China
John Wiersema, Smithsonian Institution, United States

***Correspondence:**
Carlos Magdalena
c.magdalena@kew.org
Oscar A. Pérez-Escobar
o.perez-escobar@kew.org
Alexandre K. Monro
a.monro@kew.org

† 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: *24 February 2022*
Accepted: *09 June 2022*
Published: *04 July 2022*

Citation:
Smith LT, Magdalena C, Przelomska NAS, Pérez-Escobar OA, Melgar-Gómez DG, Beck S, Negrão R, Mian S, Leitch IJ, Dodsworth S, Maurin O, Ribero-Guardia G, Salazar CD, Gutierrez-Sibauty G, Antonelli A and Monro AK (2022) Revised Species Delimitation in the Giant Water Lily Genus Victoria (Nymphaeaceae) Confirms a New Species and Has Implications for Its Conservation. Front. Plant Sci. 13:883151. doi: 10.3389/fpls.2022.883151

INTRODUCTION

Reliably documenting plant diversity is necessary to protect and sustainably benefit from it. At the heart of this lie species concepts and the practical methods used to delimit taxa. Since Darwin first linked the phenomenon of speciation to that of evolution, systematic biologists have largely conceived of species mechanistically, equating them with separately evolving lineages equivalent to branches of the 'Tree of Life' (de Queiroz, 2007; Padial and De la Riva, 2021), with the logical consequence that the basis and process of species delimitation centres on assigning individuals to a phylogenetic lineage (de Queiroz, 1988, 1999, 2007; Mayo, in press). However, others argue that lineage divergence alone is not sufficient to delimit species (Freudenstein et al., 2017; Wells et al., 2021; Lavin and Pennington, in press). For example, Lavin and Pennington provide several examples in plants of mechanisms that yield paraphyletic species. Templeton in his 'Cohesion Species Concept' instead applies explicitly evolutionary criteria to define species as, "the most inclusive group of organisms having the potential for genetic and/or demographic exchangeability" (Templeton, 1989, p. 181). The heuristic approach, which is an extension of this idea, aims to reconcile the theory and practice of species delimitation (Wells et al., 2021). This proposes species as the outcome of their constituent individuals responding to similar suites of ecological and evolutionary forces in the same way and recognizes them in practice due to congruence in properties shaped by these forces. In addition, by advocating the application of multiple categories of data, we believe that this approach to delimiting species best reflects both the quality and the breadth of observations available (Monro, in press), whilst also overcoming the limitations of lineage-based approaches (Wells et al., 2021).

Here we seek to apply a heuristic approach to the delimitation of species in the giant water lily genus *Victoria* Lindl. (Nymphaeaceae), a small and charismatic taxonomic group of short-lived perennial aquatic species distributed in the Amazonas and Chaco biogeographical regions of South America and famed for their enormous prickly leaves and massive blooms (**Figure 1**).

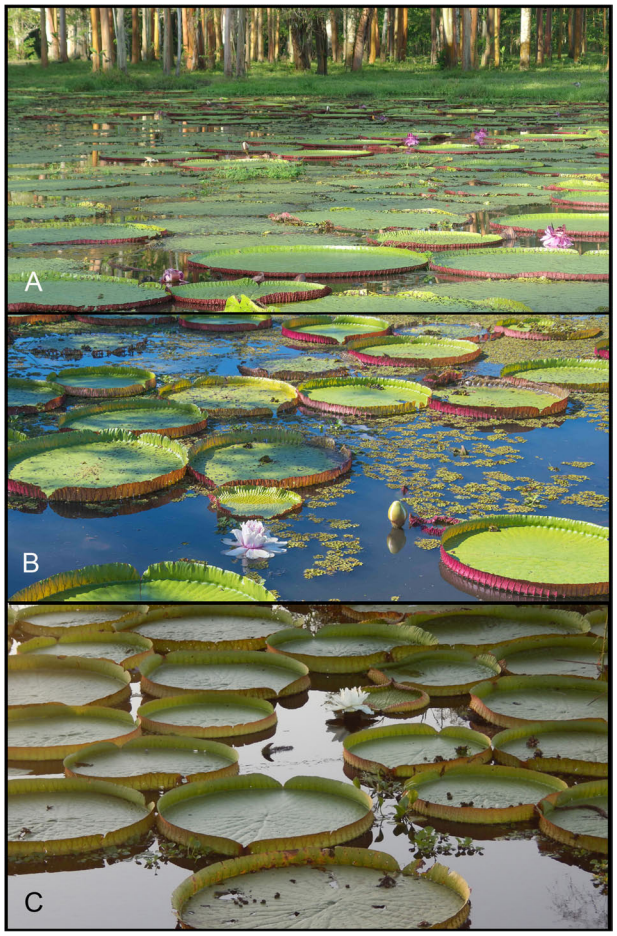

FIGURE 1 | Wild populations of **(A)** *Victoria amazonica* (Peru, Loreto), **(B)** *V. boliviana* sp. nov. (Bolivia, Beni) and **(C)** *V. cruziana* (Argentina, Chaco). (Photo Credits: **(A)** Laurel Allen, iNaturalist ID 821386. **(B)** Carlos Magdalena. **(C)** Fernanda Alarcón, iNaturalist ID. 19516477).

Indigenous Names for Giant Water Lilies

Long before the description of the taxa by Poeppig (1832), Schomburgk (1837), and d'Orbigny (1840), *Victoria* was well-known to the Indigenous Peoples of South America, featuring in indigenous narratives (Prance and Arias, 1975) and with only partially documented cultural usages.

Local names for *V. amazonica* that have been recorded include '*auapé-yaponna,*' after auapé (*Jacana jacana*), a small bird often seen running on its leaves (Schultes, 1985; Box et al., 2022). *Victoria cruziana* has been called '*yrupé,*' '*yacare yrupé,*' or '*naanók lapotó*' ('*poncho del yacaré*') (Crovetto, 2012; Scarpa and Rosso, 2014; Mereles et al., 2020).

Assigning the Genus Name *Victoria* – In Search of Patronage

In 1832, after having been noted by several botanical explorers ("Haenke in 1801 and Bonpland in 1819," and d'Orbigny in 1827; d'Orbigny, 1835, 1840), the German explorer and naturalist Eduard Friedrich Poeppig described a giant Water Lily, *Euryale amazonica*, that he had encountered in the environs of the Solimões river in Brazil that same year. Despite news of this discovery spreading quickly through Germany, it apparently reached neither Paris nor London and five years later the same species was described again, almost simultaneously, by the German botanist, Robert Schomburgk (Schomburgk, 1837) and the British botanist John Lindley (Lindley, 1837), under two different species epithets (*regina*, *regia*) and as a new genus (*Victoria*). Lindley's species epithet prevailed in usage, possibly because of the great pomp with which it was delivered and its use to lobby a new monarch:

"It is therefore not less my duty …. in distinguishing your Majesty's illustrious name, by far the most majestic species in the family of the Nymphs – one of the most noble productions of the Vegetable Kingdom – found in your Majesty's South American dominions by a gentleman [Schomburgk] traveling under the auspices of your Majesty's Government. . ." [Lindley, 1837].

The naming of the species was of great significance for the British scientific establishment as it came at a strategically important time, in the first months of the reign of Queen Victoria and at a time when several institutions were lobbying for royal patronage. Lindley's opportunistic description of *Victoria regia* not only helped the Royal Geographical Society and the Horticultural Society of London (now Royal Horticultural Society) obtain patronage from Queen Victoria (Opitz, 2013), but also contributed to the decision not to close the Royal Botanic Gardens, Kew (Hooker, 1847).

According to Opitz (2013), it was also the first signal of the prominent status that science would come to play during her reign. The description of this species therefore had a much broader impact on botanical science in the British Victorian era, at a time when it was arguably at its most influential.

Given the fashion for greenhouses and cultivation of exotic plants, it is not surprising that the cultivation of *Victoria* in Europe and North America became a symbol of social status and horticultural achievement (Holway, 2013; Aniśko, 2014), with several Botanic Gardens even constructing dedicated greenhouses for the purpose. It could also be argued that it became a symbol of the British Empire, Paxton incorporating the structure of the leaf into the architecture of Crystal Palace (Aniśko, 2014).

Further Nomenclatural History of *Victoria amazonica*

For most of the 19th and 20th centuries, the giant water lily from Amazonia was incorrectly known by the binomial *Victoria regia*. This was despite two earlier binomials for the taxon having priority, *Victoria* R. H. Schomb. for the genus name, and *amazonica* Poepp. for the species epithet.

Lindley had described *Victoria regia* [Victoria Regia: 3 (1837)] to accommodate material collected by Robert Hermann Schomburgk in Guyana. He did so amid some controversy (d'Orbigny, 1840; Opitz, 2013; Aniśko, 2014). Not only had Lindley prepared his manuscript in secrecy and against Schomburgk's instruction, but he did so a month after Schomburgk had unwittingly published his in the Athenaeum (September 9 vs. October 16), thanks to a presentation on Schomburgk's behalf by John Edward Gray to the Botanical Society (Gray, 1837). In all likelihood, Lindley's secrecy reflected his appreciation of the role it could play in securing royal patronage and he wanted his name to have priority over John Edward Gray's publication, whose preparation he was aware of. Lindley was, however, likely unaware of Poeppig's earlier description [Froriep's Not. Natur- Heilk. 25: 131 (1832)] but was soon made aware of it by a German correspondent, Weissenborn, who drew attention to the fact in the Magazine of Natural History (Weissenborn, 1837), a widely read British publication. On hearing of this earlier name, the epithet of which would have had priority over '*regia*,' he did not either respond or correct the nomenclature and it was only 10 years later that the German botanist Johann Friedrich Klotzsch published the corrected combination, *V. amazonica* (1847). Lindley's epithet, however, remained that most commonly applied to material of *V. amazonica* for over a century until Ghillean Prance's clarification of the nomenclature of the species (Prance, 1974), after which time it was slowly replaced by *V. amazonica* (Poepp.) Klotzsch.

A Second Giant Water Lily Species Described – *Victoria cruziana*, the Taxonomy of the Genus

Victoria cruziana Orb. was described as a second species by the French botanist Charles Henry Dessalines d'Orbigny [Ann. Sci. Nat., Bot., sér. 2, 13: 57 (1840)] based on material that he had collected in Corrientes, Argentina (considered Bolivia by d'Orbigny) in 1832. Since then, several additional names have been published, *Euryale bonplandia* Rojas Acosta [Cat. Hist. Nat. Corrientes: 151 (1897)], *Euryale policantha* Rojas Acosta [Cat. Hist. Nat. Corrientes: 65 (1897)], *Victoria trickeri* (Malme) Mutzek [based on *V. cruziana* f. *trickeri* Malme, Gartenwelt 29: 616 (1925)], all of which have subsequently been considered as synonyms of either *V. amazonica* or *V. cruziana*. In addition, two forms of *V. cruziana* were described by the Swedish botanist Gustaf Oskar Andersson Malme in 1907 [Acta Horti Bergiani 4(5): 12. 1907], *Victoria cruziana* f. *trickeri* Malme and *Victoria cruziana* f. *mattogrossensis* Malme.

Poeppig (1832) and Lindley (1837) disagreed over which genus the giant South American water lily should be assigned to, Poeppig assigning it to *Euryale* Salisb., a monotypic genus of large- leaved spiny water lilies from southern and eastern Asia, Lindley and Schomburgk considered it a distinct genus restricted to the neotropics. Both opinions are supported by molecular analyses which consistently recover species of *Victoria* and *Euryale* as sister to each other, either within (Löhne et al., 2007; Borsch et al., 2008), or sister to (Les et al., 1999; Borsch et al., 2008; Pellicer et al., 2013; Zhang et al., 2020) *Nymphaea*.

Ethnobotanical Significance of *Victoria* in Its Natural Range

The large seeds of *V. cruziana* are consumed as a substitute for maize (Arbo et al., 2002; Mereles et al., 2020) and the rhizomes also have recorded usage as food (Scarpa, 2009). Similarly, amongst communities inhabiting the Paraguay river in the Pantanal region of Brazil, the seeds of *V. cruziana* f. *mattogrossensis* are ground with a pestle into a starch (Bortolotto et al., 2015). The petioles also have recorded usage as food (Kinnup and Lorenzi, 2014) and a juice obtained from the roots of *V. amazonica* is a source of natural black dye, used locally to color hair (Rosa-Osman et al., 2011). Medicinal uses of *V. amazonica* include wound treatment (Schultes, 1990), and *V. cruziana* has been recorded as an anti-inflammatory and a means for combating respiratory illnesses (Hurrell et al., 2016). The total cultural, spiritual and ethnobotanical knowledge of *Victoria* discovered by Indigenous Peoples is certainly more extensive, but poorly documented in the literature.

Flower Morphology

Whilst the floral morphology of *Victoria* (**Figures 2**, **3**, **4**–**6** and **Table 1**) has been well described, there has been no

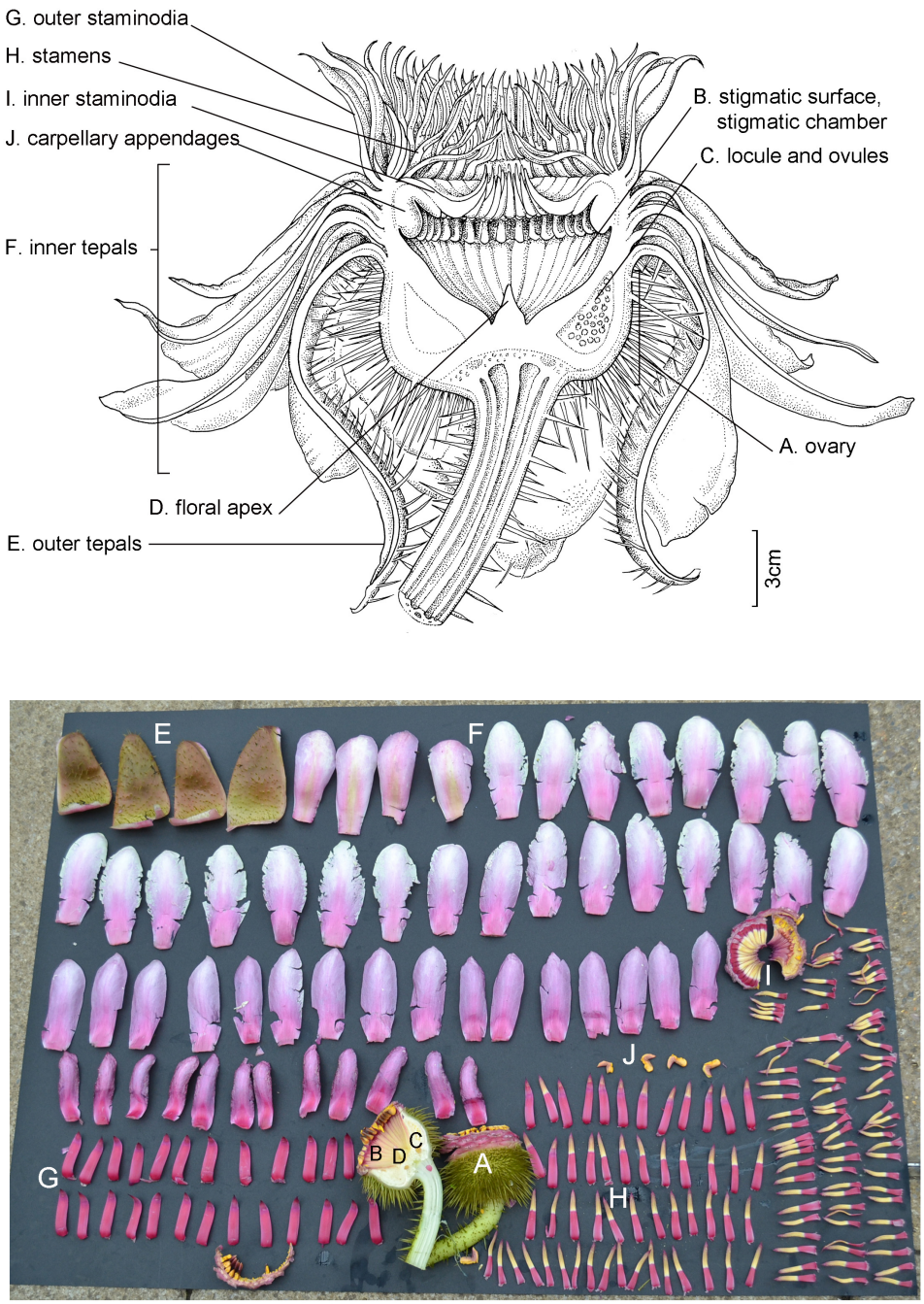

FIGURE 2 | Flower morphology and terms, using a *Victoria amazonica* second-night flower in longitudinal section for reference (above) and a fully dissected flower (below). **(A)** Ovary, **(B)** stigmatic surface and stigmatic chamber, **(C)** locule and ovules, **(D)** floral apex, **(E)** outer tepals, **(F)** inner tepals, **(G)** outer staminodia, **(H)** stamens, **(I)** inner staminodia, **(J)** carpellary appendages. Illustration and photo: Lucy T. Smith.

consensus in the terminology applied. The flowers have an inferior ovary, comprising 25–40 radially arranged syncarpous carpels (**Figure 2A**). The upper part of the ovary is a concave, papillose stigmatic surface, which is divided into raised segments (**Figure 2B**). Each segment corresponds to the roof of the locule below and bears a dorsal longitudinal slit through which pollen reaches the locules. Each locule contains 8–28 ovules, attached parietally to both sides of the locule's walls (**Figure 2C**). At the center of the stigmatic surface is a column of residual stelar tissue, also known as the floral apex (Schneider, 1976; **Figure 2D**). There has been some inconsistency in the description of the perianth segments. Following Warner et al. (2008), we are referring to all perianth segments as tepals. From the apex of the ovary's external rim, four rigid fleshy outer tepals arise (**Figure 2E**; Warner et al.,

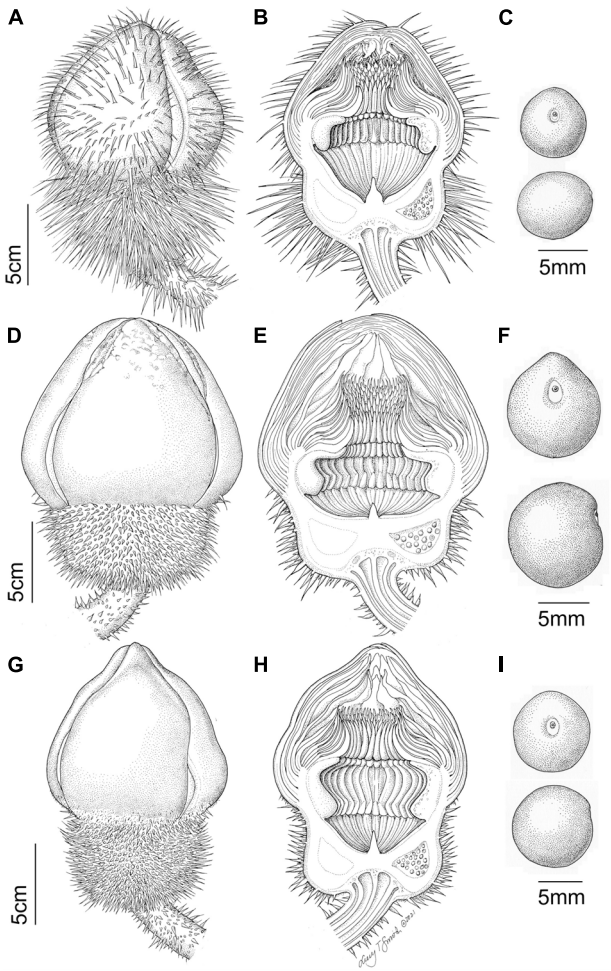

FIGURE 3 | Line drawing comparing flower and seed morphology in: *Victoria amazonica* **(A–C)**, *V. boliviana* sp. nov. **(D–F)**, and *V. cruziana* **(G–I)**; flower bud whole and in longitudinal section showing abaxial surface of outer tepals, bud profile, stigmatic chamber, carpellary appendage and prickle morphology; seeds. (cultivated at RBG Kew) Illustration: Lucy T Smith.

2008). From the tepals, helically arranged series of petaloid inner tepals follow (**Figure 2F**). Adjacent to these, and moving inwards, rigid, thick, outer staminodia arise, which are presented in a whorl-like arrangement of one or two whorls (**Figure 2G**). In bud and on the first night opening, the outer staminodia (along with the stamens and inner staminodia which follow) are sigmoid, and arch strongly over the stigmatic surface (**Figures 3B,E,H**). After these are sigmoid stamens (**Figure 2H**), borne in two or three whorls, pressed tightly against the outer staminodia in bud and during the first night of opening. The stamens are followed by sigmoid inner staminodia in one or two whorls (**Figure 2I**). The inner staminodia (previously classed as paracarpels: Prance and Arias, 1975) are partially adnate to the upper portion of the carpellary appendages, which lie beneath them. These L- shaped carpellary appendages (**Figure 2J**) are aligned with the locules below, and correspond with them in number. The lower portions of the carpellary appendages are abaxially fused to the rim-like extension of the stigmatic surface which is also fused with the basal tissue of the inner tepal bases.

The cavity enclosed by these parts is referred to as the stigmatic chamber (**Figure 2B**). In bud and during the first night of flowering, the apices of the outer staminodia, stamens and inner staminodia are pressed tightly into each other to form an entrance tunnel (**Figures 3B,E,H**). This tunnel remains intact in bud and during the first night of opening, providing entry to the beetle pollinator, but is absent on the second night of opening, when all the outer staminodia, and most of the stamens, reflex to varying degrees (**Figure 2**). The inner staminodia, however, do not reflex, but instead fall further downwards, thus blocking the entrance to the stigmatic chamber. Flower size, the number of locules and the number of all parts (apart from the outer tepals, which always number four) are variable both between individual plants and on different flowers produced by the same plant.

Pollination and Dispersal Biology

Victoria flower buds develop underwater and emerge above the surface when ready to bloom. Each flower opens over two consecutive nights, changing form and color dramatically in-between. These form and color changes reflect their role in pollination, which is to trap pollinating beetles (Schomburgk, 1837; Prance and Arias, 1975) of the Cyclocephalini tribe (Scarabaeidae) (Prance and Arias, 1975). Floral morphology of the fossil *Microvictoria*, suggests that this mode of pollination was already established in the Cretaceous (Gandolfo et al., 2004). The carpellary appendages are believed to produce scent attractants and nutritional rewards (Prance and Arias, 1975; Zini et al., 2019) and with the warming of the flower (thermogenesis) at night (Planchon, 1850, 1851; Knoch, 1899; Decker, 1936; Prance and Arias, 1975; Lamprecht et al., 2002) serve as both an attractant and stimulant for the pollinators (Seymour and Mathews, 2006). The stigmatic surface remains receptive for the two nights during which the flower blooms (Archangeli, 1908). Despite the above investment in cross-pollination, Prance and Arias (1975) demonstrated that self-pollinated flowers were still capable of setting seed. Furthermore, seed produced in cultivation as a result of selfing is viable (Magdalena, personal observation). There have been few published pollination studies of *V. cruziana* and only one record of the putative new species from the Mamoré river basin in Bolivia (Magdalena, personal observation). After pollination, the fruit forms below the water surface (Prance and Arias, 1975). The seeds are covered by a mucilaginous tissue that has been proposed to represent an aril, are buoyant for a few days and released as the fruit decomposes (Prance and Arias, 1975). Prance and Arias (1975) suggest that seeds of *V. amazonica*, produced at the end of the wet season, are dispersed over long distances because of the annual flooding of much of its habitat and this may also be the case for *V. cruziana*. Although there is no evidence of endochory in the genus, this should not be excluded as the dispersal biology of the species remains very poorly studied.

Ecology and Biogeography

Victoria occurs in white-water and occasionally black-water (Byrne, 2008) leas and igapos of the Amazonian and Paraná river basins (Rosa-Osman et al., 2011) at depths of up to

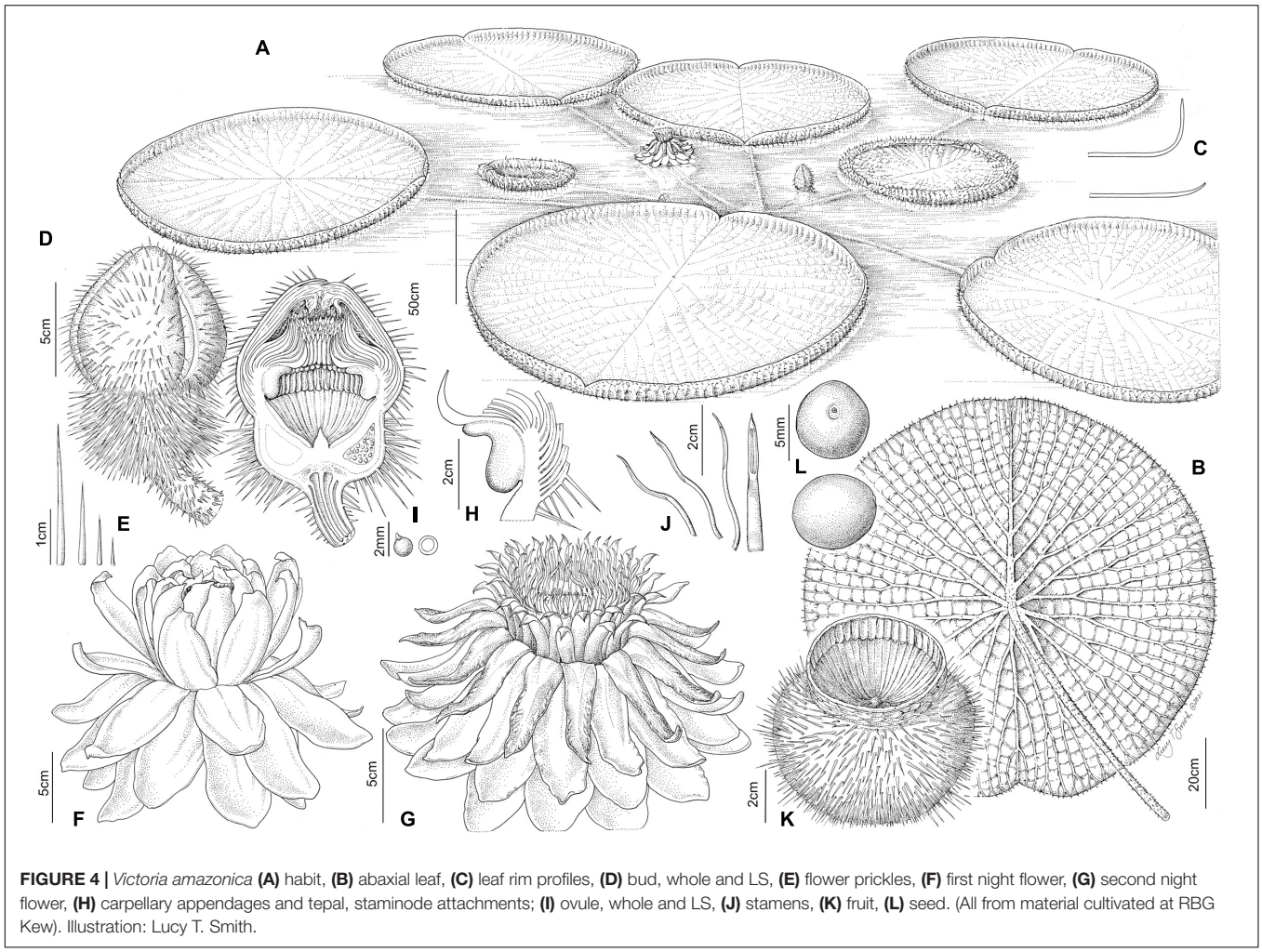

FIGURE 4 | *Victoria amazonica* **(A)** habit, **(B)** abaxial leaf, **(C)** leaf rim profiles, **(D)** bud, whole and LS, **(E)** flower prickles, **(F)** first night flower, **(G)** second night flower, **(H)** carpellary appendages and tepal, staminode attachments; **(I)** ovule, whole and LS, **(J)** stamens, **(K)** fruit, **(L)** seed. (All from material cultivated at RBG Kew). Illustration: Lucy T. Smith.

5.25 m (*V. amazonica*, Prance and Arias, 1975). *Victoria* spp. are commonly classified as annuals but Prance and Arias (1975) propose that, in the case of *V. amazonica* at least, this life cycle likely reflects a constraint posed by the dramatic changes in water level associated with flooding and drought, which characterize the wet and dry seasons across the range of this species. In a stable horticultural environment, *V. amazonica* and *Victoria* 'Longwood' hybrids can thrive for several years. They should, therefore, be considered to be short-lived perennial species. The seeds of *V. amazonica* are desiccation intolerant (Rosa-Osman et al., 2011). Seedlings develop very quickly in the river mud forming mature plants in three to five months. Development is quicker in *Victoria cruziana* compared to *V. amazonica* (Kit Knotts et al., personal observation). This is possibly a reflection of the shorter and more predictable growing season in the temperate biome of *V. cruziana.* Senescence is triggered by the detachment of the rhizome from the riverbed, or desiccation as the river level drops (Aniśko, 2014).

Taxonomic Challenges

Taxonomic treatment of *Victoria* has been hampered by several factors. Foremost is the fact that the type collections of the two currently recognized species (World Checklist of Vascular Plants, 2022) have been lost or destroyed, making it challenging to unambiguously name material and thus delimit species. Poeppig's collection(s) at Naturhistorisches Museum Wien and University of Leipzig were likely destroyed during WW2 and d'Orbigny's s spirit collection disappeared from the Paris museum for reasons unknown. In addition, d'Orbigny diagnosed *V. cruziana* against material of *Victoria* morphotype *'boliviana'* that he had mistakenly considered to be *V. amazonica* and in doing so established misleading species-limits at a time when very few collections were available for study. The above issues were compounded by the fact that *Victoria* is notoriously difficult to make herbarium specimens from, being big, fleshy, covered in prickles and prone to rotting in the dryer. This likely explains why there are relatively few herbarium specimens of wild collected material: only 97 of *V. amazonica* and 18 of *V. cruziana* (The Global Biodiversity Information Facility [GBIF], 2022), despite the species' broad distributions.

These reasons could explain why the genus attracted relatively little taxonomic attention in the 20th Century, during which the most notable contributions were made by Malme (1907) and Prance and Arias (1975). In the 21st Century, renewed interest

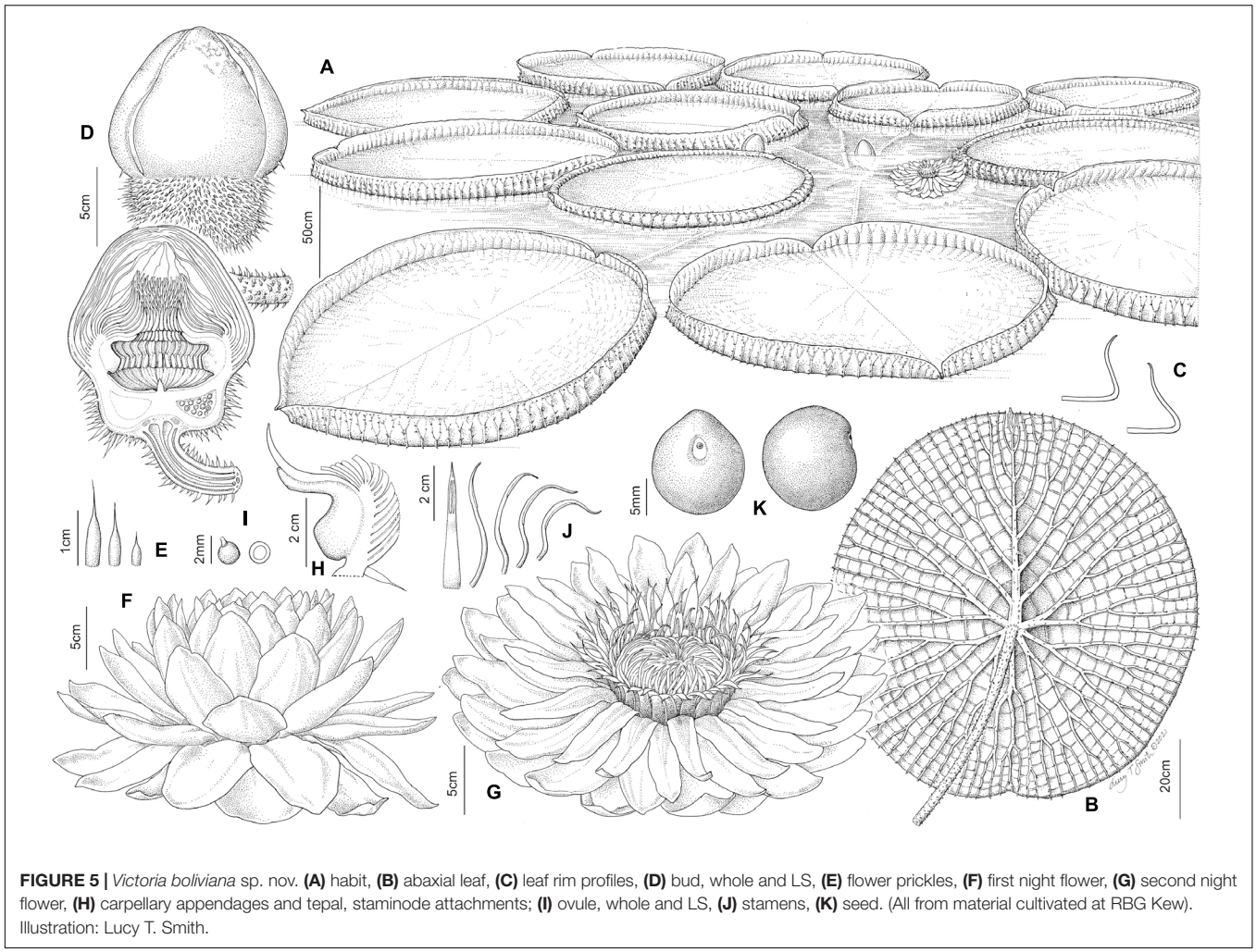

FIGURE 5 | *Victoria boliviana* sp. nov. **(A)** habit, **(B)** abaxial leaf, **(C)** leaf rim profiles, **(D)** bud, whole and LS, **(E)** flower prickles, **(F)** first night flower, **(G)** second night flower, **(H)** carpellary appendages and tepal, staminode attachments; **(I)** ovule, whole and LS, **(J)** stamens, **(K)** seed. (All from material cultivated at RBG Kew). Illustration: Lucy T. Smith.

was, associated with research for the Flora of Brazil. Notably Pellegrini's treatment for the Flora of Brazil (2020) that recognizes the two species, and de Lima et al. (2021), which recognize a single variable species (*V. amazonica*).

Significantly, much of our knowledge regarding the natural history of *Victoria* has been obtained from material growing in cultivation and compiled by horticulturalists in the additional literature or worldwide web[1] and it was observations by C. Magdalena which first suggested the existence of additional species and the need for a re-evaluation of the genus.

The aims of this study were to (1) revisit species delimitation in *Victoria* and to do so through an iterative process, using morphological observations to establish species hypotheses and suites of observations to test these, and in doing so (2) reveal the principle evolutionary lineages, their age, permeability and biogeography, (3) identify diagnostic morphological and DNA sequence characters for the species, and use the above to (4) revise the nomenclature, descriptions, distribution, and conservation status of the species.

[1]http://victoria-adventure.org/

MATERIALS AND METHODS

We applied a heuristic species concept (Wells et al., 2021) in which morphological, field and horticultural observations were used iteratively to develop initial hypotheses of species limits that we further tested using rigorous phylogenomic and population genomic analyses. This study was an Anglo-Bolivian collaboration instigated simultaneously by both parties with the goal of conducting research into *Victoria* in an equitable manner (McAlvay et al., 2021).

Morphological Evaluation of Putative Species

Taxon Sampling

Morphological observations were made from as many gatherings and images of plants as possible. We examined 110 sheets of 58 herbarium collections both physically and digitally (via Reflora and JSTOR Plants) held at BM, COR, HGCS, IAN, JBRJ, K, LPB, MO, NY, P, SI, SPF, UBCB, and US (abbreviations according to Thiers, 2016). These were supplemented with 175 'research grade' geolocated field images from iNaturalist, images publicly available

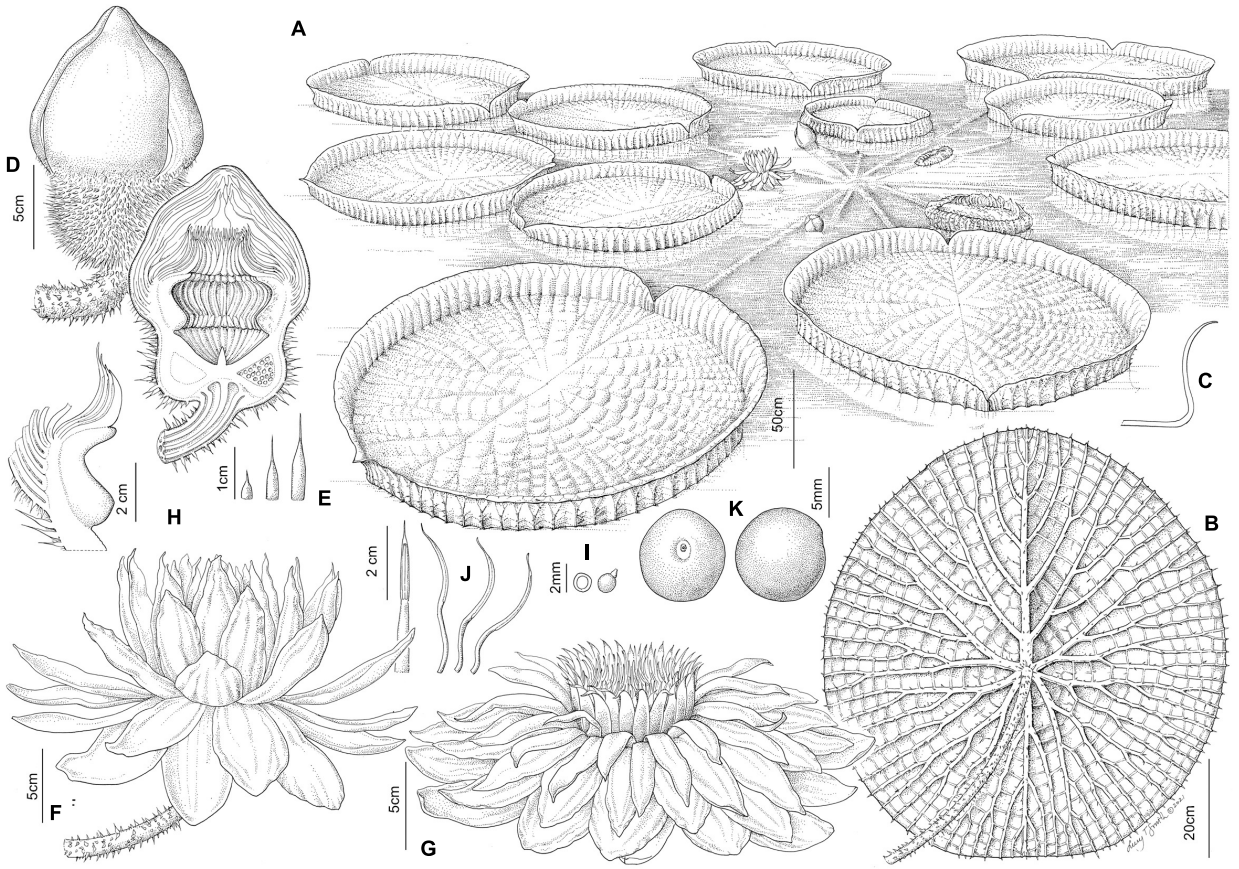

FIGURE 6 | *Victoria cruziana* **(A)** habit, **(B)** abaxial leaf, **(C)** leaf rim profiles, **(D)** bud, whole and LS, **(E)** flower prickles, **(F)** first night flower, **(G)** second night flower, **(H)** carpellary appendages and tepal, staminode attachments; **(I)** ovule, whole and LS, **(J)** stamens, **(K)** seed. (**A–E, G–K**, cultivated at RBG Kew, **F** cult. Denver Botanic Gardens). Illustration: Lucy T. Smith.

on Instagram, Facebook, Googlesearch, and living collections at K. A subset of the herbarium specimens was used as the source of DNA for the genomic analyses (**Supplementary Table 1**). This subset was designed with the aim of including representatives of at least three gatherings per morphotype and principal river basin (Amazon, Essequibo, Paraná), including those from or close to the type localities of *Victoria amazonica*, *V. cruziana* and *V. cruziana* f. *mattogrossensis*. These included a type of element of *Victoria cruziana* (d'Orbigny s.n., P02048598).

Morphological Observations

The selection of the morphological characters recorded was based on our field and horticultural observations and experience of examining herbarium material, characters used in previous studies (notably Malme, 1907) and those which could be observed from herbarium, living collections or research quality geo-referenced photographic images on iNaturalist. The herbarium and iNaturalist datasets complemented each other well since each favored a particular set of characters (**Table 1**). For example, herbarium collections were a good source of observations of prickle morphology, number and distribution, carpellary appendages, tepal pigmentation, trichomes, seed size and shape. iNaturalist images were a better source of observations of gross morphological characters such as the shape and height of leaf rims and flower color. Where close-up photos were included in iNaturalist records, it was also possible to examine prickle morphology, number, and distribution.

Twenty-seven morphological characters were documented for living collections at RBG Kew (**Table 1**), encompassing observations of every part of the plant. These were also used to calculate rim heights as a proportion of leaf length (**Tables 1**, **2**). Nineteen morphological characters were documented for herbarium specimens (**Table 1** and **Supplementary Table 4**), encompassing: leaf indument (rim, lamina), presence or absence and distribution of prickles on outer tepals and their morphology, number and distribution, tepal number, carpellary appendage size and shape, ovary indument and prickle morphology, fruit prickle morphology, color of innermost tepals in bud, and seed size. Due to the size of *Victoria* leaves, the size of a herbarium sheet and the flattening of leaf structures, specimens were not a source of observations on leaf size or rim height. Twenty-one morphological observations were documented from iNaturalist images (**Table 1** and **Supplementary Table 5**) encompassing adaxial and abaxial leaf

TABLE 1 | Source of contrasting morphological characters and their states from herbarium, horticultural, or field image data.

Character	Herbarium specimens	iNaturalist	Kew cultivated
(1) Leaf blade color adaxial (green, bronze, maroon)		X	X
(2) Leaf blade color abaxial (green, dark green/blue, yellow, maroon)		X	X
(3) Leaf rim shape (absent/present. Curving gently or perpendicular/recurved over adaxial surface at base/sigmoid & flared at top)		X	X
(4) Leaf rim height proportionate to length (absent, low, moderate, high, and as percentage of length)		X	X
(5) Leaf rim color abaxial (pale green/"white," green, green tinged pale maroon, deep maroon)		X	X
(6) Hairs on abaxial leaf (size and number of segments, total hair length, density)	X		
(7) Shape of bud (apex convex/concave toward tip)	X	X	X
(8) Prickles on outer tepals abaxial, presence or absence	X	X	X
(9) Prickles on outer tepals abaxial, distribution (covering or partially covering, where on outer tepal)	X	X	X
(10) Prickles on outer tepals abaxial, number	X	X	X
(11) Prickles on outer tepals abaxial, spacing (regular or irregular)	X	X	X
(12) Prickles on outer tepals abaxial, shape (smoothy tapering to sharp point or abruptly tapering to sharp point)	X	X	X
(13) Prickles on outer tepals abaxial, range of sizes and actual sizes	X	X	X
(14) Hairs on outer tepals (visible to naked eye or not)	X	X	X
(15) Prickles on ovary (present on all species), shape (smoothly tapering to sharp point or abruptly tapering to sharp point)	X	X	X
(16) Prickles on ovary (present on all species), number of sizes, and actual sizes	X		X
(17) Color of innermost tepals in bud and on first-night opening (white/dark or maroon)	X	X	X
(18) Shape and depth of stigmatic chamber (deep/shallow, oblong/triangular/rounded)		X	X
(19) Color of open first-night flower, inner tepals (white, white with maroon innermost tepals)		X	X
(20) Texture of inner tepals on first-night opening (smooth/crinkled)	X	X	X
(21) Color of open second-night flower inner tepals (white, pale pink, dark pink, maroon)	X	X	X
(22) Carpellary appendages, shape of lower part (fixed, arising flat from surface/free, curved)	X	X	X
(23) Carpellary appendages, size of upper part and lower parts and their relationship (smaller than/greater than/equal to equal to)	X	X	X
(24) Size of ovules	X		X
(25) Number of ovules per locule			X
(26) Number of seeds per fruit			X
(27) Seed size	X		X
(28) Seed shape (globose, ellipsoid, raphe distinct or not)	X		X

color, rim shape and color, bud, first- and second-night flower color, and occasionally floral characters such as the distribution, morphology and number of outer tepal prickles and the color of innermost tepals in bud.

Geographical Observations

Geographical observations were used both to map the distribution of putative *Victoria* species and to undertake extinction risk assessments. Localities were taken from the labels of herbarium collections and the metadata of iNaturalist records and recorded as decimal coordinates (**Supplementary Table 2**).

For collections where there was no coordinate data but where precise locality information was given, GoogleEarth was used to estimate latitudes and longitudes. We also reviewed social media accounts of water lily enthusiasts using tags for *Victoria* (Facebook, Instagram, Googlesearch). Whilst useful, social media posts were not considered as a reliable source of geographical coordinates for Victoria populations; unlike iNaturalist, these platforms do not curate spatial data and posts can be removed or edited at any time. Social media posts were deemed unsuitable for use in the calculation of extinction threat assessments but valuable in providing an indication of hitherto undocumented populations that could then be confirmed by directly contacting the posters and confirming the locality using Googlearth. Because *Victoria* grows in open stretches of riverbank, it is possible to recognize *Victoria* populations in Google Earth images, due to their distinctive leaf outline and size, and contrast with a body of water.

Extinction Risk Assessments

Extinction risk assessments were undertaken using IUCN Red List Categories and Criteria of Threatened Species (Hereafter IUCN Red List) version 3.1 (IUCN, 2001; IUCN Standards and Petitions Committee, 2019). Calculations of the extent of occurrence (EOO) and area of occupancy (AOO) were undertaken using the online conservation assessment tool GeoCAT (Bachman et al., 2011). The estimated AOO was calculated using a cell width of 2 km as recommended by IUCN and the estimated EOO was calculated based on the minimum convex polygon (IUCN Standards and Petitions Committee, 2019). Due to the availability of a relatively small number of herbarium specimens

TABLE 2 | Morphological character states used to delimit morphospecies of *Victoria*.

Character	*V. amazonica*	*V. boliviana*	*V. cruziana*	*V. c.* f. *mattogrossensis taxon incertum*
Rim shape of mature leaf in cross-section	Curved at base, perpendicular Not flared at apex	Strongly recurved over adaxial surface Curling inwards or flared at apex	Recurved over adaxial surface at base Flared at apex, sigmoid	Strongly recurved over adaxial surface Curling inwards or flared at apex
Rim height	Absent, or low to moderate, 4–7% leaf length, higher only in congested areas	Moderate, 5–7% leaf length	Moderate to high, 8–10% leaf length	Moderate, unknown % leaf length
Rim color	Maroon, occasionally green	Maroon/red, or pale green	Green, or tinged pale maroon	Maroon/red
Leaf trichome length (where present)*	0.3 – 12 mm	1.2 – 3 mm	1 – 3 mm	?
Leaf trichome segment number*	3 – 12	6 –15	10-15	?
Bud shape*	Convex at apex	Convex at apex	Concave just before apex	Convex at apex
Ovary prickles shape	Smoothly tapering to a point	Abruptly tapering tapering to a point	Abruptly tapering to a point	Abruptly tapering to a point
Ovary prickle size (dried)	2 – 21 mm	1 – 10 mm	1 – 22 mm	2 – 15 mm
Ovary trichome length (where present)*	0.1 – 0.4 mm	NA	0.1 – 12 mm	??
Ovule number*	20–28	8–14	20–25	??
Ovule size (fresh) (dried)	 1.5 mm 0.5 – 1.5 mm	 2–2.5 mm 2–2.5 mm	 1.5 – 1.8 mm 1.2 —1.5 mm	 – –
Abaxial outer tepal color	brown/maroon	Green or maroon	Green or maroon	Green or maroon
Abaxial outer tepal prickles	Present	Absent or present	Absent or present	Present
Abaxial outer tepal prickle number*	55 – 330	0 – 10	0 – 100	300 – 1000+
Abaxial outer tepal prickle shape	Smoothly tapering to a point	Abruptly tapering to a point	Abruptly tapering to a point	Abruptly tapering to a point
Abaxial outer tepal prickle distribution*	Covering entire surface	Covering entire surface	Covering basal third only	Covering entire surface
Abaxial outer tepal prickle arrangement	Regularly to irregularly spaced	Irregularly spaced	Regularly to irregularly spaced	Regularly spaced
Abaxial outer tepal prickle (dried)	1 – 14 mm Two – three sizes, two sizes predominant	2.5 – 5 mm Three sizes	1 –7 mm One – four sizes	1 – 4 mm Three sizes, two predominant
Abaxial outer tepal trichome length (where present)*	0.1–0.2 mm	NA	0.1 – 1 mm	??
Color of innermost tepals in bud and on first night opening*	Dark red/maroon	White	White	White
Proportionate lengths of upper and lower arms of L-shaped carpellary appendages*	Upper part equal to or shorter than lower part	Upper part greater than lower part	Upper part smaller than lower part	Upper part smaller than lower part
Shape of lower arm carpellary appendage at base*	Rounded, hanging free, auriculate	Straight, partly free	Straight, attached	Straight, partly free
Seed shape	Ellipsoid	Globose	Globose	Globose
Seed raphe*	Raphe faintly visible	Raphe prominent	Raphe faintly visible	Raphe prominent
Seed dimensions*	7–8 × 9–10 mm	12–13 × 16–17 mm	8–9 × 9–10 mm	6–10 × ca 10 mm

** denotes novel character in the table.*

we calculated and contrasted a maximum and minimum range. The maximum range included potential habitat within the range, whilst the minimum range was limited to confirmed observations, either from herbarium specimens or iNaturalist posts. Unverified images were defined as "presence uncertain" and excluded from the minimum estimate but included for the maximum range. The extinction risk assessments undertaken here will be uploaded to the IUCN Species Information Service (SIS) in 2022 after completion of the official peer reviewed process and official submission to the IUCN Red List.

Genomic Evaluation of Putative Species

Taxon Sampling

Leaf tissue was sampled from 18 specimens obtained from both herbarium collections (n = 12, K, MO, HGCS, and P) and

living collections [n = 6, Adelaide Botanical Garden, Australia (AD), K, Santa Cruz Botanic Garden, Bolivia, Royal Botanic Garden, Kew (K)]. Samples from the living collections included three individuals of the putative new species (seeds from Santa Cruz Botanic Garden) as well as an outgroup [*Nymphaea ampla* (Salisb.) DC.] (**Supplementary Table 1**). Samples from living collections and the field are hereafter referred to as "fresh" samples. These tissue samples were stored in silica gel prior to DNA extraction. All specimens were used in compliance with loan agreements of the source biological collections (K, MO, P, and HGCS).

DNA Extraction, Library Preparation and Sequencing

A total of 20–40 mg leaf tissue was weighed out and pulverized using a SPEX® sample prep tissue homogenizer (SPEX Inc, Metuchen, NJ, United States). DNA was extracted using CTAB and isopropanol (Doyle and Doyle, 1987) and cleaned using a 2x ratio of AMPure XP beads (Beckman Coulter, Brea, CA, United States). DNA libraries were prepared using NEB Next Ultra II Library Prep Kits according to the manufacturer's protocol (with half volume reactions) and with NEBNext Multiplex Oligos for Illumina (New England Biolabs, Ipswich, MA, United States) amplified with 9–11 PCR cycles (fresh samples) or 11–15 PCR cycles (herbarium samples). Yield and fragment size distribution were estimated using a Quantus fluorometer (Promega, Madison, WI, United States) and a 4200 TapeStation system (Agilent Technologies, Santa Clara, CA, United States) respectively. Sequencing of DNA libraries was carried out on an Illumina NovoSeq platform with a paired end 150 bp configuration, by GeneWiz® (South Plainfield, NJ, United States).

Generating Transcriptome-Based *Victoria* Nuclear Reference Reads

Since no genome assembly is currently available for genus *Victoria*, we used transcriptome reads to create a *Victoria* genomic reference for read mapping. The published transcriptomic data was obtained from a *V. cruziana* sample (SRX6884057) (Zhang et al., 2020) sequenced on an Illumina platform. We trimmed adaptor sequence from the reads using Trimmomatic v. 0.39 (Bolger et al., 2014), with sliding window trimming, cutting once the average quality across 4 bases fell below a PHRED score of 20 and requiring a minimum length of 30 bp. We then carried out a *de novo* assembly of the reads using Trinity v.2.8.5 (Grabherr et al., 2011) with default settings. We estimated transcript abundance (where a minimum threshold value acts as a proxy for 'real' genes) using the alignment- free method *salmon*. Subsequently, we filtered the raw assembly for transcripts of low expression with a normalized TPM (transcripts per million) matrix, where transcripts with an expression level < 1 TPM for any given sample (an expression level of at least one could be of biological relevance) and retained only the most expressed isoform of each transcript. This resulted in retention of 74,088/152,932 (48.45%) transcripts. We then used CD-HIT (Fu et al., 2012) to cluster all of the transcript sequences and retain only one read from any clusters of similar sequences, where the identity threshold was set to 0.95. This step filtered out 172 (<0.4%) of transcripts. Finally, we removed all transcripts of length < 350 bp (an assumed minimum insert size, given that Illumina sequencing of 150 bp paired-end reads was conducted). This resulted in a *Victoria* genomic reference set of 38,703 transcript sequences. To ensure that our population genetic analyses were exclusively derived from nuclear SNPs and not from organellar or fungal/bacterial DNA sequences, we conducted a remote blast search of the newly assembled transcriptome against the entire NCBI nucleotide database, using the *blastn* software of the NCBI tools (Camacho et al., 2009), an e-search value of 0.001 and keeping a maximum of five hits per queried sequence. We discovered that 820 (2.1%) and 612 (1.6%) of the transcripts respectively matched chloroplast and mitochondrial sequences, and thus were removed from subsequent analysis. In addition, ∼0.85% of the total proportion of blasted transcriptomes matched bacterial or fungal DNA sequences, indicating that the presence of contaminant reads in the assembled transcriptome is negligible. Our filtered *Victoria* reference set derived from the assembly totaled 37,470 transcripts. Finally, to assess completeness of the *de novo* assembly, we used BUSCO v. 5.3.2 (Simão et al., 2015), applying the lineage dataset chlorophyta_odb10 [constituting 16 genomes and 1519 benchmarking universal single-copy ortholog (BUSCO) genes].

Processing of High-Throughput DNA Sequence Data and Alignment to Transcripts

We trimmed the raw read data using AdapterRemoval v2.3.2 (Schubert et al., 2016) with the 'collapse' option to maximize retention of shorter reads, a consideration based on our dataset having a large proportion of herbarium specimens (Latorre et al., 2020). We aligned the trimmed reads to the transcriptomic reference set of reads using bwa v 0.7.17 (Li and Durbin, 2009), with the *mem* algorithm (suited to long reads and seeds alignments with exact matches) for the samples from fresh material and *aln* for the herbarium material (suited to short reads and allows for mismatches). We retained reads with a minimum mapping quality of 20 and of a minimum length of 25 bp (herbarium samples) and 30 bp (fresh samples) and removed PCR duplicates using the function *rmdup* of the software samtools v.1.7 (Li et al., 2009). Endogenous content (the proportion of *Victoria* DNA sequence compared to exogenous reads) was calculated by comparing totals of mapped reads (before PCR duplicate removal) to totals of trimmed reads. Mean sequencing depth was calculated along the entirety of the aligned sequence for each sample.

Population Genomic Analysis of Nuclear Data

Given the shallow phylogenomic scale under investigation in this study and due to the high prevalence of inter-specific hybridization characterizing the family *Nymphaeaceae* (Borsch et al., 2014; Robson et al., 2016), and the potential of inter-specific hybridization to interfere with inference of the species trees in flowering plant taxa (Pirie, 2015; Morales-Briones et al., 2021; Pérez-Escobar et al., 2021a,b), we applied a population genomic approach. We excluded the *Nymphaea* outgroup from this analysis, leaving our set of 16 *Victoria* samples. Due to

the low average depth of sequencing afforded by our genome skimming approach (see *Results*), we estimated genotypes using a genotype likelihood (GL) method in ANGSD v.0.933 (Korneliussen et al., 2014). In this approach, genotype likelihoods were scored inferring major and minor alleles and retaining sites with *p*-value of at least 1e-6 and a minimum mapping quality of 30. We used these genotype likelihoods to carry out a principal component analysis (PCA) with PCAngsd (Meisner and Albrechtsen, 2018), with default settings and a maximum of 10,000 iterations. Due to the very small sample size, we chose a stringent threshold for genotype missingness across all samples – a maximum of 1 missing individual before rejection of the site (*-minInd* set to 15) and minor allele frequency (*maf*) thresholds of 0.1 and 0.2 were applied. Minor alleles can have a disproportionately large effect on population structure inference (Schmidt et al., 2021); singletons and doubletons will be very common given the very restricted sample size here, thus we would expect the results from the *maf* = 0.2 filtering run to more accurately represent the true genomic structure. The *maf* = 0.1 iteration was performed for comparison, as a dataset comprising more genotyped sites in total.

Phylogenomic Analysis of Plastid Data

For the construction of a plastid phylogeny, we utilized a published chloroplast genome, of *V. cruziana* (Gruenstaeudl et al., 2017) available on the NCBI repository (NC_035632) as a reference genome. We aligned our trimmed reads to this reference and filtered them using bwa v 0.7.17 (Li and Durbin, 2009), using the same settings as above (see section "*Processing of high-throughput DNA sequence data and alignment to transcripts*"). We then used ANGSD to generate pseudohaploid (where diploid genomic data is simplified into a single consensus sequence) consensus sequences from the aligned reads, setting a minimum sequencing depth threshold of 10 and a minimum base quality score of 30. Using the 15 samples in which genotyping completeness was > 99.8%, along with *Euryale ferox* as an outgroup, we computed the maximum likelihood (ML) tree using the software RAxML v.8.2.12 (Stamatakis, 2014), with a GTR substitution model, the GAMMA model of rate heterogeneity and 500 bootstrap replicates. The genome for *Euryale ferox*, the tropical Asian water lily (NC_037719.1) (He et al., 2018) was sourced from the NCBI repository.

Molecular Dating Analysis

To elucidate the absolute times of divergence amongst populations of *Victoria*, we relied on the implementation of molecular clocks and multispecies coalescent (MSC) models in the program StarBEAST2 v.2.5 (Ogilvie et al., 2017), on the same whole plastid genome data produced to compute a ML phylogeny (see section "*Phylogenomic analysis of plastid data*"). This approach enables the estimation of calibrated species trees using population sampling information while considering topological gene tree incongruence such as the one derived from incomplete lineage sorting (ILS) of gene flow (Ogilvie et al., 2017). One partition representing the entire whole chloroplast genome and a total of 158,992 sites (of which 390 were informative) was used as input, containing linear sequences of three individuals representing populations of *V. 'boliviana'* morphotype, four individuals representing *V. cruziana*, eight individuals representing *V. amazonica*, one individual of *Euryale ferox* and one individual of *Nymphaea ampla*, the latter two employed as outgroups. Following the times of divergence obtained by Zhang et al. (2020) for Nymphaeales, to calibrate our plastid phylogeny, we relied on two secondary calibration points applied to: (a) the root of the tree representing the divergence of *Victoria* and *Euryale* from *N. ampla*, set to 75 Ma, (b) to the MRCA of *Euryale* and *Victoria*, set to 36 Ma; both secondary calibration points were set to a normal prior distribution and a standard deviation of 1. We modelled the substitution rates with a GTR substitution model and rate heterogeneity among sites with a four-categories Gamma distribution in conjunction with a relaxed log-normal molecular clock. The molecular clock was informed using a uniform prior distribution for the mean rate, ranging from 1.0e-5 to 0.001, which represents a range of plastid substitution rates reported for different land plant lineages (Gaut et al., 1992). A ploidy level of "1" (option "Y or mitochondrial") was indicated in the program, as recommended for plastid datasets (Drummond and Bouckaert, 2015). Lastly, a coalescent constant population tree model with a mean population size of 1.0 and a non-informative prior of 1/X was chosen, following (Drummond and Bouckaert, 2015) whenever a mixture of population-level sampling is involved. We executed 100 million generations in StarBEAST2, sampling every 5000 states and ensuring that all parameters reached convergence as evidenced by effective sample sizes > 200.

Comparative Genomics of *Victoria* Plastomes

To further investigate the genomic properties of the proposed new species (specifically: consistent variation in the form of point mutations, indels or structural variation), we *de novo* assembled the plastid genomes of *V. cruziana* (n = 2), *V. amazonica* (n = 1), and *V. 'boliviana'* morphotype (n = 2) and created whole genome alignments. We took the raw reads from samples: NPNY23 (*V. cruziana*), NPNY21 (*V. cruziana*), NPNY14 (*V. amazonica*), NPNY24 (*V. 'boliviana'* morphotype), NPNY26 (*V. 'boliviana'* morphotype), trimmed these using stringent settings in Trimmomatic v.0.39; retaining only reads at least 50 bp long and removing bases with a Phred quality score below 30. These trimmed reads were the raw material for plastome assembly we subsequently performed with GetOrganelle (Jin et al., 2020), using default settings. The resulting complete genomes were aligned using Mauve v.2.3.1 (Darling et al., 2004), as implemented in the platform Geneious v.8.1.9 (Kearse et al., 2012). In this alignment, apart from the assembled genomes, we included plastid genomes from the NCBI repository: *V. cruziana* (NC_035632) as well as the plastid genome of *Euryale ferox*. Whole genome alignment was conducted using the mauveAligner algorithm with the following options: 'full alignment,' 'extend local collinear blocks (LCB)' and 'automatically calculate minimum LCB score.' Genomes were functionally annotated using the software GeSeq (Tillich et al., 2017) of the Chlorobox toolkit[2] and the following

[2]https://chlorobox.mpimp-golm.mpg.de/geseq.html

parameters: a protein search identity value of 25, rRNA, tRNA and DNA search identity of 85, and the annotated plastid genomes of *E. ferox* (NC_037719.1) and *V. cruziana* as reference (NC_035632). The resulting annotated LCBs were scanned manually to detect indels and point mutations unique to the *V. 'boliviana'* morphotype sequences. Finally, each LCB was analyzed in DnaSP v.6 (Rozas et al., 2017) to compute the variant sites between the three *Victoria* species, parsimony-informative sites and point mutations unique to *V. 'boliviana'* morphotype.

Genome Size Estimation

Nuclear DNA contents were estimated by propidium iodide flow cytometry using fresh leaf material. Around 1 cm^2 matured leaf tissue from the specimen was co-chopped with the internal standard [*Petroselinum crispum* (Mill) Nyman ex A. E. Hill 'Champion Moss Curled,' 1C = 2171.16 Mb (Obermayer et al., 2002)] using a new razor blade in 1 ml of General Purpose Buffer supplemented with 3% PVP (GPB) (Loureiro et al., 2007). A further 1 ml of GPB was added to the sample and the contents gently mixed in the petri dish. The sample was then passed through a 30 μm nylon filter. The homogenate was stained with 100 μl propidium iodide (1 mg/ml) and incubated on ice for 10 min. Two samples were prepared from the same individual and three replicates of each were run, recording up to 1,000 nuclei per fluorescence peak using a Sysmex CyFlow Space (Sysmex Europe GmbH, Norderstedt, Germany) flow cytometer fitted with a 100 mW green solid state laser.

The resulting histograms were analyzed with the WindowsTM-based FlowMax software (v. 2.9 2014, Sysmex GmbH) and the average of each sample was used to estimate genome size.

Chromosome Count

A chromosome count was obtained from a single individual of *V. 'boliviana'* morphotype. growing in the RBG Kew Living Collection (accession number x2018-659) using the conventional root squash method to observe mitotic chromosomes (Pellicer et al., 2007). Briefly, actively growing root tip meristems were collected and pre-treated with aqueous colchicine [0.05% (v/v)] for 4 h, then transferred to freshly made Carnoy's fixative [3:1 (v/v) absolute ethanol and acetic acid] for 24 h at ∼21°C. Root tips were then transferred to 70% (v/v) ethanol and stored at –20°C until used. Before squashing, root tips were hydrolyzed in 1M HCl at 65°C for 8 min, transferred to 2% aceto-orcein and stored at 4°C overnight. Each root tip was placed on a microscope slide and squashed under a 22 × 30 coverslip in a drop of 4.5% acetic acid and analyzed under a Leitz Laborlux D phase-contrast microscope (Ernst Leitz Wetzlar GMBH, Germany).

RESULTS

Morphological Observations

We scored morphological character states from 58 herbarium collection specimens and 175 iNaturalist observations, giving a total of 233 individuals (see **Supplementary Tables 4, 5**). The congruence of morphological character states to our four morpho species was evaluated by eye, resulting in diagnostic characters being selected (see **Table 2** and **Figures 4**, **5**, **7**, Key to the species): leaf rim morphology (shape in cross-section, height, color), flower bud apex shape, size and shape of the stigmatic chamber, outer tepal prickle morphology, number and distribution, inner tepal color in bud, carpellary appendage morphology and attachment, seed shape, size and presence / absence of a raphe. These were represented by 25 characters, 13 of which are novel.

Geographical Observations

We recorded 209 geographical observations assignable to morphospecies (**Figure 3** and **Supplementary Table 2**), 175 of these correspond to iNaturalist records (indicated by * in **Supplementary Table 2**), and 44 to biological collections in herbaria. In several cases, iNaturalist records extended the distribution of *Victoria* delimited by herbarium collections (**Figure 3**). For example, in the case of the *cruziana* morphotype, herbarium collections indicate a southernmost limit of the –30.37°S latitude, whilst iNaturalist increased this to –32.92°S. Social media posts provided additional probable distributions for *Victoria* not documented in herbaria or iNaturalist. For example, images of *V. amazonica* from Irinida in Colombia, close to the Irinida and Guaviare rivers, both of which drain into the Orinoco river; and images of *V. cruziana* from the Esteras del Ibera, Argentina.

Our review of herbarium and iNaturalist records (**Supplementary Table 2** and **Figure 8**) and of social media images suggest that *Victoria* is absent from a number of river systems that form part of the Amazon, Essequibo and Paraná river basins (**Figure 8** and **Supplementary Table 3**). The most notable of these being its absence from central eastern and north western Amazonian Brazil and, despite being documented in Colombian tributaries of the Orinoco river, its apparent absence from Venezuela and Ecuador.

Morphospecies Correlate Well With Geography

Using the geographical observations (**Supplementary Table 1**) we mapped our four morphospecies across South America (**Figure 8**). This shows strong congruence between morphospecies and geographical location. Based on our dataset, *amazonica* and '*boliviana*' morphotypes are restricted to mutually exclusive portions of the Amazon river basin, the '*mattogrossensis taxon incertum*' morphotype is restricted to the Pantanal (Paraná river basins), and *cruziana* is restricted to the lower portion of the Paraguay river and Paraná. Despite the presence of both *cruziana* and the '*mattogrossensis taxon incertum*' morphotype on the Paraguay river, occurrence records suggest a large geographical separation between both.

Distribution

Victoria is largely restricted to temperate and tropical Southern Hemisphere South America, not occurring further than 4.2°N and 32.9°S. Whilst it has an extensive temperate distribution

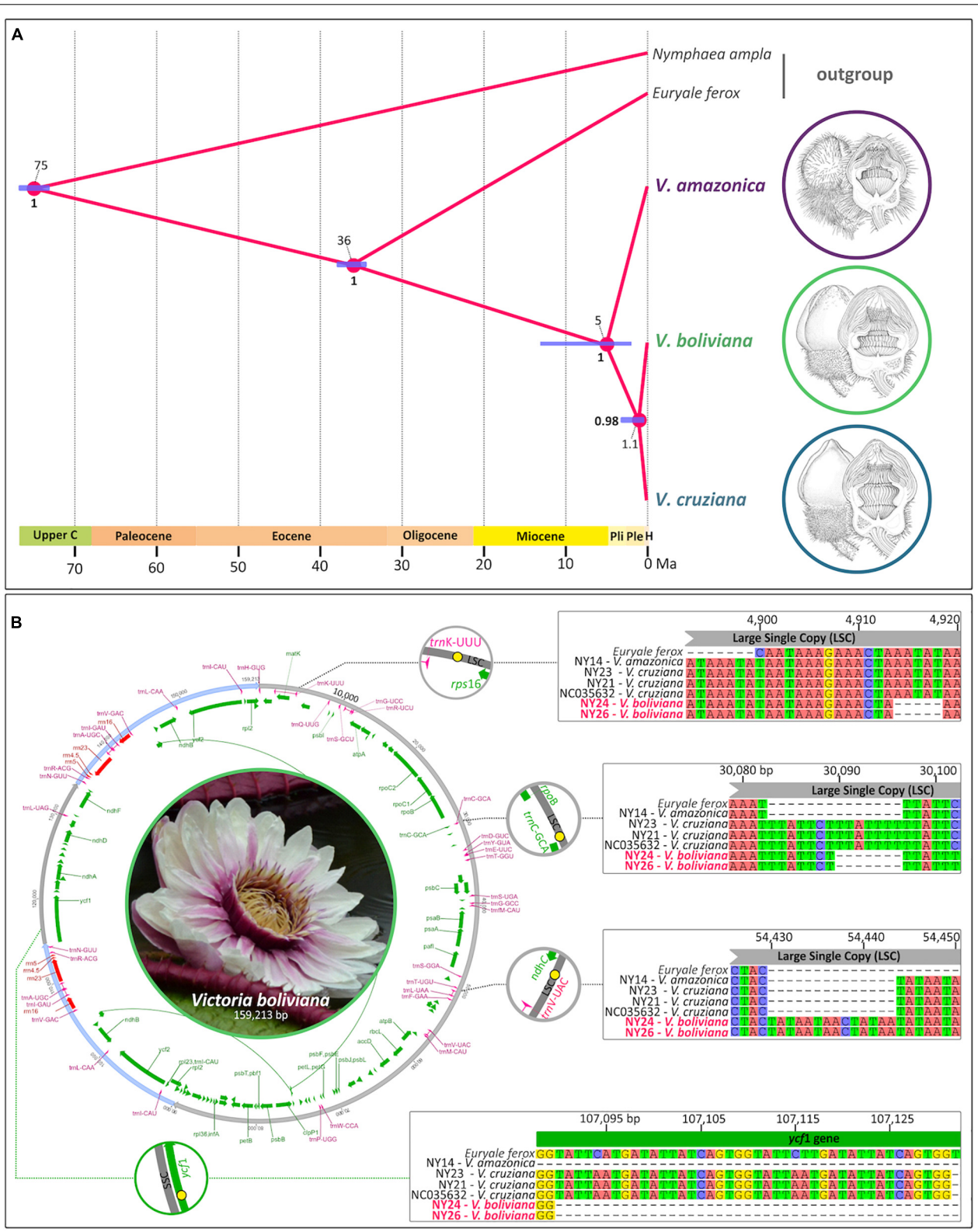

FIGURE 7 | (A) Chronogram tree from StarBEAST2 representing molecular dating analysis of splits between *V. amazonica, V. cruziana,* and *V. boliviana* with confidence intervals highlighted. **(B)** Visual representation of plastid structure of *V. boliviana* assembled using GetOrganelle, with genomic locations of diagnostic characters indicated. Details of polymorphisms (indels) – in the context of multi plastid alignments – displayed within insets.

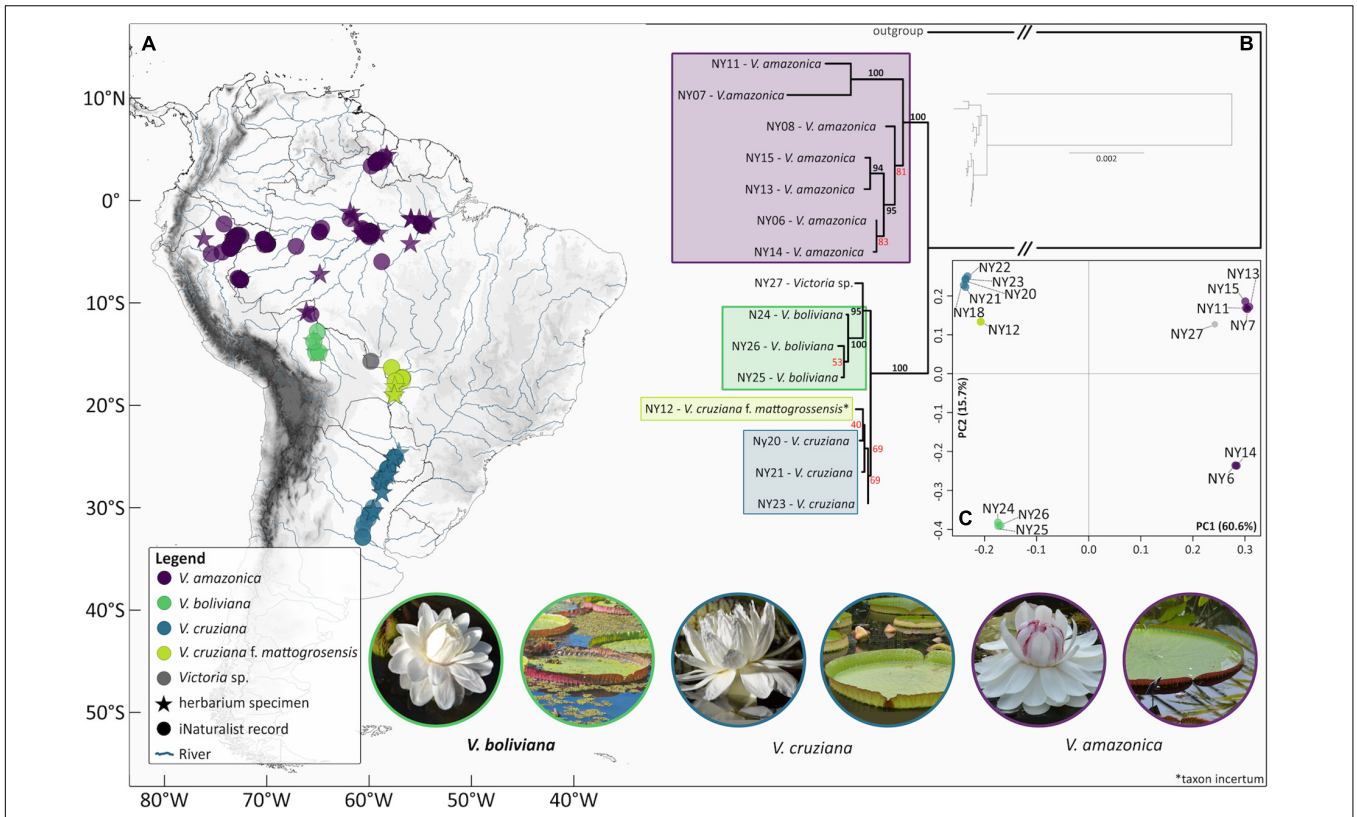

FIGURE 8 | Geographical context with population genomic and phylogenomic results. **(A)** Geographical coordinates of all *Victoria* samples used in analyses, where symbols are color-coded according to morphotype/species and symbol type denotes data source. **(B)** Phylogenomic representation of relationships between plastomes of 15 *Victoria* samples, with *Nymphaea ampla* as the outgroup and constructed using RAxML. **(C)** Principal component analysis of *Victoria* nuclear dataset in 16 individuals based on 229,281 genotype likelihoods obtained after skimming for a minor allele frequency (*maf*) threshold of 0.2.

in the southern hemisphere, it does not occur in the temperate Northern Hemisphere outside of cultivation.

V. amazonica is known from Colombia, Brazil, Guyana, Peru, and Bolivia where it is restricted to the river basins of the Amazon, Guaviare river (a tributary of the Orinoco) and Essequibo rivers. In the Amazon river basin *V. amazonica* occurs in most of the major tributaries except for the Xingu and Madeira rivers. *V. cruziana* is the only temperate species and is restricted to the Paraná River and its tributaries. The *V. 'boliviana'* morphotype appears to be endemic to the Llanos de Moxos in Bolivia with all the records concentrated in the Mamoré river basin. A photographic record from Rurrenabaque, however, suggests that it may also be present in the neighboring Beni river.

Genomic Evaluation of Putative Species

DNA Sequencing and Mapping of Reads to Plastid and Transcript Datasets

Library insert size, including adaptors, was on average 195 bp for herbarium specimens and 385 bp for fresh collection samples. A total of 3.9-12.5 million raw reads were generated for each sample (**Supplementary Table 7**). Mapping reads to the plastid genome resulted in an average read depth of 74x (herbarium material) and 1716x (fresh material) (**Supplementary Table 8**), where the proportion of missing sites was <0.7% for all samples included in the final analyses. As for reads mapped to the set of 37,470 transcripts that constitute our substitute for a nuclear reference genome, their average depth totaled 4x and 29x for herbarium and fresh material respectively (**Supplementary Table 9**). The length of quality-filtered sequence data aligned to the transcripts totaled up to 36 Mbp in herbarium samples and 43 Mbp in fresh samples. Our BUSCO analysis revealed that our assembled transcriptome captured 1287 (84.7%) of complete BUSCO genes, where 23 (1.5%) were fragmented and 209 (13.8%) missing.

Population Genomic and Phylogenomic Analysis of Nuclear Data

For the analysis performed using GLs, the number of positions obtained for each run with different filtering parameters was: 436,329 after *maf* = 0.1 filtering (**Supplementary Figure 1**) and 229,281 after *maf* = 0.2 filtering (**Figure 8**). The resulting PCAs suggest that the most informative axis of variation derived from nuclear genomic sequence data of these *Victoria* samples distinguishes *V. amazonica* from *V. cruziana* with *V. boliviana* sp. nov. (40.9 and 60.6% of variation along the 1st PC in respective runs), whereas a smaller proportion of the variation (28.0 and 15.7% in respective runs) collectively describes the segregation of *V. cruziana* from *V. boliviana* sp. nov. as well as the majority

of *V. amazonica* from the *V. amazonica* samples from Guyana (NPNY6 and NPNY14). In the *maf* = 0.2 iteration of the analysis, the separation of Guyana *V. amazonica* from the remainder of samples is less pronounced than for the *maf* = 0.1 filtering run. Overall, the spread of samples implies more population genomic structuring within *V. amazonica* compared to that within *V. cruziana.* Regarding the *V. c.* f. *mattogrossensis taxon incertum* sample, based on these nuclear variant sites, it shows an expected affinity of NPNY12 to *cruziana*. The unidentified sample (NPNY27) shows close affinity to the main *V. amazonica* cluster. A very similar pattern of genetic clustering was produced with the *maf* = 0.1 filtering option (**Supplementary Figure 1**).

Phylogenomic Analysis of Plastid Data

Our phylogenomic tree (**Figure 8**) based on alignments of the entire plastid sequence (with *Nymphaea ampla* and *Euryale ferox* as outgroups) shows strong support for the monophyly of *V. amazonica* samples (100% of BS replicates for this bipartition). Within this clade, there is additionally robust support (94%) for two Brazilian samples (NPNY13 and 15) in the Rio Solimões/Manaus area as well as 100% support for two other Brazilian samples (NPNY7 and 11) provenanced to Municipio de Oriximiná and Ipixuna respectively. The adjacent bipartition (100% support) subtends sister bipartitions including: all of the *cruziana* samples (including *V. c.* f. *mattogrossensis taxon incertum*) (69% support) and all three *V. boliviana* sp. nov. samples along with the unassigned sample NPNY27 (95% support), where the three *V. boliviana* sp. nov. samples have 100% support. Notably, branch lengths are shorter within the *V. cruziana*, *V. c.* f. *mattogrossensis taxon incertum*, *V. boliviana* sp. nov. clade than within the *V. amazonica* clade. The topology derived from the absolute age estimation analysis conducted in StarBEAST2 was in strong agreement with the ML individual level phylogeny. Here, *V. amazonica* was placed as sister to *V. cruziana* and *V. boliviana* sp. nov. with maximum support. The sister relationship of *V. cruziana* and *V. boliviana* sp. nov. was also recovered with strong support (Posterior Probability = 0.98, for HPD intervals, see **Figure 7**). The analysis further revealed that populations of V. *cruziana* and *V. boliviana* sp. nov. diverged in the Pleistocene, 1.1. Ma, whereas *V. amazonica* branched out from the MRCA of *V. cruziana* and *V. boliviana* sp. nov. 5 Ma (**Figure 7**).

Comparative Genomics of Plastomes Amongst *Victoria* spp.

Four locally collinear blocks (LCBs) were identified by the Mauve alignment. Amongst our set of plastid genomes derived from four species (*E. ferox, V. amazonica*, *V. cruziana,* and *V. boliviana* sp. nov.), no genomic-scale rearrangements were detected, including in the inverted repeat (IR) regions. We detected 12 insertions and deletions differentiating intraspecific *Victoria* genomes, ranging from 4 to 105 bp in size. These included two indels specific to *V. boliviana* sp. nov: a 14 bp insertion in the large single copy region (LSC), two deletions in the LSC (5 and 7 bp in length), the former between *trnK* and *rps16*, and the latter adjacent to *trnC*, and a 42 bp deletion in the coding sequence (CDS) of gene *ycf1*, situated within the small single copy region (SSC) (**Table 3**).

Furthermore, a 4 bp transversion unique to *V. boliviana* sp. nov. was found in the LSC. For the three species, seven sample alignment here, DnaSP analysis revealed 182 polymorphic sites, where 17 sites classified as parsimony informative and 8 SNPs constituted alleles private to *V. boliviana* sp. nov. (**Supplementary Table 6**).

Genome Size Estimation and Chromosome Count

The flow cytometric analysis of *V. boliviana* sp. nov. resulted in high resolution flow histograms with the 2C peaks of both the sample and internal calibration standard having low coefficients of variation (CV%) (mean of 2.77 for samples, and 2.99 for calibration standard). Based on the means of the sample G1 and calibration standard G1 peaks, *V. boliviana* sp. nov. has an estimated genome size of 1C = 4.24 pg (**Supplementary Figure 2**). In addition, a chromosome count of $2n = 2x = 24$ was obtained for the same accession of *V. boliviana* sp. nov (**Supplementary Figure 3**). The count is identical to that previously determined for *V. cruziana* and different from that for *V. amazonica* of $2n = 2x = 20$.

DISCUSSION

Heuristic Delimitation of *Victoria* Species

Through the application of a heuristic multidisciplinary approach, we provide revised species delimitations and diagnoses for *Victoria*. This has resulted in the recognition of a species new to science: *V. boliviana* Magdalena and L. T. Smith and highlighted the morphological distinctness of *V. cruziana* forma *mattogrossensis taxon incertum*. Species delimitation and nomenclatural stability in *Victoria* have until now been hampered by the loss of the original type material that served to fix the species name as well as a paucity of biological collections. This resulted in disagreement over the number of species recognized (Pellegrini, 2020; de Lima et al., 2021), the application of an incorrect name for *Victoria amazonica* for most of the 19th and 20th centuries (Prance, 1974) and a failure to recognize taxa in this iconic genus. Nevertheless, the recent selection of a neotype for *V. amazonica* and a lectotype for *V. cruziana* (de Lima et al., 2021) has helped to underpin nomenclatural stability and anchor species delimitation with respect to morphology. Here, we sought to delimit *Victoria* species, through the development of a heuristic and iterative approach – one which integrates field, horticultural, morphological and genomic observations and analyses. In the primary iteration of this investigation, we used geographical observations to integrate morphological observations from biological (herbarium) collections with those from field observations (iNaturalist, horticultural observations). In doing so, we were able to recognize discrete morphological units within a heuristic species concept that focuses on cohesion rather than divergence. These formed *prima facie* null hypotheses that we tested using genomic observations. Based on this approach, we recognize three discrete units at the rank of species, and provide strong justification for further research into a fourth taxon of unknown rank. Our morphological species delimitation differs from that of de Lima et al. (2021) who propose considering all populations of *Victoria* as a single species,

TABLE 3 | Plastid polymorphisms (other than point mutations) diagnostic for *V. boliviana* sp. nov. Alignment blocks are arbitrary units computed by Mauve v.2.3.1.

Alignment block	Polymorphism type	*V. boliviana*	*V. cruziana*	*V. amazonica*	Starting position (alignment block)	Plastid region	Proximal genes (flanking unless specified)
LCB1	Deletion	5 bp			4,914	LSC	*trnK rps16*
LCB1	Deletion	7 bp		14 bp	30,090 (30,083 *amazonica*)	LSC	*trnC*
LCB2	Insertion	14 bp			3,499	LSC	*ndhC trnV*
LCB3	Deletion	42 bp		105 bp	17,916 (17,853 *amazonica*)	SSC	*ycf1* (intragenic)
LCB2	4 bp transversion	AAAA	TTTT	TTT-	34,612	LSC	*rpl36 rps11*

Victoria boliviana sp. nov. deletions and insertion are relative to V. cruziana, where two of the deletions overlap with longer respective deletions in V. amazonica. LSC, large single copy region; SSC, small single copy region.

V. amazonica. Their conclusion was based on the interpretation of the resulting large variation in the morphological characters observed in their single species as being the product of its broad spatial distribution and aquatic habit.

Overcoming the Challenge of Small Numbers and Poorly Preserved Biological Collections

Past studies (e.g., de Lima et al., 2021) have been limited by the small number of herbarium collections available to study and their state of preservation, both of which are likely a product of the large size and fleshy nature of the plants. We overcame this through the use of high-resolution specimen scans available online, supplemented by "research quality" geo-referenced iNaturalist field images and observations of horticultural material. Using iNaturalist and digitized herbarium specimens we were able to incorporate 233 collections of *Victoria*, a significant increase on previous studies based on herbarium specimens alone (de Lima et al., 2021). iNaturalist images and observations allowed for characters usually lost in herbarium such as leaf rim morphology, the color of leaf blades, flower bud apex shape, the distribution and shape of prickles on the outer tepals, and the color and shape of tepals (**Table 1**).

The geo-location of *Victoria* lily pads enabled us to use these records to undertake assessments of extinction threat and so greatly increase the accuracy of those assessments. Whilst we did not use social media accounts as a source of georeferenced localities, we did use them to identify potential gaps in our knowledge of *Victoria*'s distribution. Instagram, which is image-based, was particularly useful in indicating the presence of *Victoria cruziana* in the Esteras del Ibera wetlands (Argentina), suggesting that it occupies a broader swathe of south eastern South America. Instagram also suggested the presence of *V. amazonica* in the Orinoco, and the cultivation of *V. cruziana* f. *mattogrossensis taxon incertum* under the names *V. amazonica* or *V. regia* in Brazil.

Diagnostic Morphological Characters for Delimiting *Victoria* Species

We identified the morphology of the leaf rim, flower prickles, the stigmatic chamber, carpellary appendage shape and size, and the seeds, as phylogenetically informative diagnostic characters in *Victoria*. Of these, we are the first to propose stigmatic chamber, carpellary appendage and seed morphology. This was surprising as these characters are all readily observable and prominent features that would be obvious to an experienced observer. A possible explanation may be that there was confusion over species delimitation stemming from d'Orbigny's (1840) mistaken diagnosis of *V. cruziana* against *V. boliviana* sp. nov., and not *V. amazonica*, as he and subsequent authors supposed. In the absence of type material or an independent class of observations this would have been difficult to resolve.

The prickles on the abaxial surface of the outer tepals have not been proposed as diagnostic in *Victoria* since Malme (1907). Based on our observations, *Victoria amazonica* always has prickles which taper smoothly to a point at their apex (**Figure 4**) and are distributed relatively evenly over the entire tepal surface. By contrast, in *V. cruziana* (**Figure 6**) the prickles are absent or relatively sparsely distributed, but, where present, they taper abruptly to a point and are distributed over only the basal third of the tepal, whilst in *V. boliviana* sp. nov. (**Figure 5**) prickles are usually absent but, if present, may be found anywhere on the tepal surface and taper abruptly to a point. We speculate that prickles play a defensive role and protect the relatively nutrient rich contents of the bud and the protein-rich beetles trapped within at anthesis but are unable to account for the variation between species in the genus given the incomplete and small number of studies focusing on *Victoria* biology or autecology.

We also discovered that the size and shape of the stigmatic chamber, and of the carpellary appendages which surround and heat it, differs between species (**Table 2**). As above this suggests a link to pollination but again, given that the pollination-biology of *Victoria cruziana* and *V. boliviana* sp. nov. is very poorly known, it is only possible to speculate that the size and shape of the stigmatic chamber may be responding to differences in pollinator type, size or number, whilst differences in the size and disposition of the carpellary appendages may be products of the need to accommodate a different size of stigmatic chamber, or to produce greater or lesser amounts of heat in relation to ambient temperatures or pollinator preferences.

We found observable differences between the seeds of all three species with respect to their shape and size, and of the prominence of the raphe. *Victoria amazonica* has ellipsoid seeds compared to the globose seeds of *V. cruziana* and *V. boliviana* sp. nov. *V. boliviana* sp. nov. has relatively large seeds with a prominent raphe, compared to *V. cruziana* and *V. amazonica*. The significance of seed shape is unclear but may be related

to dispersal through the gut of an unknown disperser, an ellipsoid seed being easier to pass than a globose one. Whilst no evidence of endochory has been found, given the lack of research it should not be excluded. Seed size has also been associated with the establishment depth of *Nymphaea* (Jacobs and Hellquist, 2011) suggesting that smaller-seeded species establish in shallower water. This would concur with the observations of previous authors (Prance and Arias, 1975; Rosa-Osman et al., 2011, Magdalena, personal observation) and suggest that *V. boliviana* sp. nov. establishes itself at greater water depths than *V. amazonica* and *V. cruziana*. Finally, similarly to *V. cruziana*, *V. boliviana* sp. nov. develops more rapidly than *V. amazonica* (Magdalena, personal observation.). As above, these aspects remain to be further studied.

Cohesion and Distinctness Characterizing Genomic Datasets Support *Victoria* Species Hypothesis

We explored complementary concepts of genomic distinctness (in the form of genetic clusters) and divergence of evolutionary lineages in order to test whether the identified morphotypes were corroborated by molecular evidence. At this shallow taxonomic scale, a PCA has more power to highlight distinctness in nuclear genomic data than phylogenomic methods; the latter can be confounded by ILS (Lissambou et al., 2019). Even with our genome skimming approach, given the size of the *Victoria* nuclear genome [1C of 4.66 (*V. amazonica*), 4.10 (*V. cruziana*) and 4.24 (*V. boliviana* sp. nov.)], which is at least double the size of most Nymphaeales species profiled using flow cytometry (Pellicer et al., 2013), the capacity to retrieve appropriate nuclear genes at high coverage for phylogenomic inference was limited. Our response was to retrieve genotypes by mapping to curated transcriptomic data and to summarize this genetic variation using a dimensionality-reduction method. The tight respective clusters of *V. cruziana* and *V. boliviana* sp. nov. on the PCA and their degree of separation along the 2nd axis of variation serves to demonstrate their genetic distinctness. Importantly, this distinctness could be due to geographical isolation alone, which is why such results must always be assessed against other lines of evidence, such as morphological differences. Our PCA additionally demonstrates a greater degree of genetic structuring and variation within *V. amazonica* compared to *V. boliviana* sp. nov. or *V. cruziana*. This is concordant with the broader geographical spread of *V. amazonica* in northern South America, though we also note the larger sample size of *V. amazonica* available for this analysis. Furthermore, we would expect to see more continuity of the *V. amazonica* cluster on the 2nd PC, linking the two samples from Guyana (NPNY6 and NPNY14) and the samples from Brazil, had we been able to genotype a more broadly geographically sampled set of accessions. Implementation of a population genetic framework using high throughput sequencing (HTS) datasets to support molecular-based species delimitation is not a widespread practice, but is gaining some traction in studies of plants. For example, where they occur in sympatry (Ikabanga et al., 2017) or where taxonomic incongruity is prevalent within the genus (Rodríguez-Rodríguez et al., 2018; Rutherford et al., 2018). Our study demonstrates a novel way to apply this approach – in the absence of a reference genome, but utilizing an available transcriptome.

The dataset of mapped full plastid genomes suggests a similar conclusion in terms of delimiting separate evolutionary units of *Victoria*. This is presented as a strongly supported monophyly of respective clades containing species *amazonica*, *boliviana* sp. nov. and *cruziana*. The longer branch lengths of samples within the *amazonica* clade also suggests ancient differentiation. Our molecular dating suggests that plastid populations in Victoria diverged ∼5 Ma, with the time of divergence of *V. boliviana* sp. nov. and *V. cruziana* set to have occurred as recently as 1.1 Ma.

An unresolved question is the degree to which hybridisation and introgression might have been involved in the evolution of *V. boliviana* sp. nov. By revealing that this species has a chromosome count identical to that of *V. cruziana*, a hybrid origin cannot be easily supported. Even though the genome size of *V. boliviana* sp. nov is intermediate between that of *V. cruziana* and *V. amazonica*, additional molecular processes such as repeat amplification and chromosome rearrangements have almost certainly been involved its genome evolution, especially given its divergence time from *V. cruziana*.

The Plastid as a Source of Molecular Characters for Species Diagnosis

Finally, we supplement these lines of evidence derived from different cellular compartments with genomic features identified in the assembled chloroplasts that we propose to be diagnostic to the new species. The absence of large-scale genomic rearrangements in our plastid genomes was not surprising given that gene order in the chloroplasts of Nymphaeales has been found to be conserved (Gruenstaeudl et al., 2017). The three indels unique to *V. boliviana* sp. nov., could be used to support the molecular identification of *Victoria* specimens. The longest indel was found in the *ycf1* gene (105 bp in *V. amazonica* and 42 bp in *V. boliviana* sp. nov.). *Ycf1* is a large housekeeping gene (Drescher et al., 2000) involved in photosystem biogenesis (Yang et al., 2016). Due to its high variability (Dong et al., 2015), *ycf1* has been highly utilized in phylogenetics (Neubig et al., 2009; Dastpak et al., 2018). Here, we shed light on the utility of this gene as a potential tool for DNA barcoding at shallow phylogenetic scales. One application could be in genome skimming studies, where retrieval of full chloroplast genomes is routine. A low-cost alternative is a simple PCR of this genomic region, where, due to length differences, gel electrophoresis could be used for species identification. Diagnoses of new taxa that incorporate DNA-based characters are not common, but can be more useful than lineage-based diagnoses, especially when applied to known or type specimens. They are also an unbiased means of species delimitation (Renner, 2016). A small number of previous studies have used diagnostic molecular characters e.g., nucleotides at certain positions of *matK* and *trnL-trnF* regions in *Buxus* spp. (Buxaceae; Gutiérrez et al., 2011) and of *nhdF* and the *ITS* region in *Brunfelsia* (Solanaceae; Filipowicz and Renner, 2012). Length variation associated with indels has been previously applied as a

form of diagnostic genetic variation – for example to the study of *Abies* (Xiang et al., 2018). Our approach has extended this to examination of the whole plastome.

Suggestions for Future Research Priorities in *Victoria*

Ensuring that species of *Victoria* remain for future generations requires that risks of extinction can be accurately evaluated and monitored. In the case of *Victoria* this requires greater knowledge of the species' natural history and autecology, specifically their pollination and dispersal biology – against which potential threats can be evaluated – and the extent and fluctuation of population sizes.

This is important as it enables threats to the viability of populations to be evaluated. Knowledge of both is at best superficial and based on a small number of field observations of *V. amazonica* and *V. cruziana* (Schomburgk, 1837; Archangeli, 1908; Prance and Arias, 1975; Hanagarth, 1993). The dispersal biology of *Victoria* is poorly known and based largely on speculation rather than observation of tested hypotheses. There is also no literature on the size, fluctuation and connectivity of populations of the three species. A viable approach for doing so would be to use publicly available time-stamped remote-sensed data, such as Google Earth image layers to monitor populations given that the lily pads can be seen in higher resolution images.

Our small molecular sample set recovered genetic structure within *V. amazonica* (**Figure 8**) and we would argue for greater sampling of the genus, especially for the edges of its Amazonian range (Colombia, Guyana, Peru, Venezuela) and in the vicinity of the Pantanal. Additionally, since morphological observations suggest a fourth species (*V. c.* f. *mattogrossensis taxon incertum*), strategic sampling for genomic work may result in the molecular support for it at the rank of species.

Establishing the status of *V. c.* f. *mattogrossensis taxon incertum* should be seen as a high priority. Should it be evaluated as a distinct species, it would be one of the most vulnerable to extinction, having the smallest range and occupying a region that has been impacted by extreme drought during the last decade (Marengo et al., 2021).

Finally, a larger sample set would allow for investigation of the barriers to dispersal within the genus and extent of gene flow between the different populations of *Victoria* species.

Understanding the latter would also require the generation of more extensive nuclear molecular datasets and a more contiguous genome of reference. Construction of a nuclear phylogenomic framework would enable the computation of introgression tests based on patterns of allele sharing between taxa (e.g., D statistics; Durand et al., 2011) and permit a more nuanced investigation of the evolutionary history of these aquatic plant species, including investigation of forces driving speciation of *V. boliviana* sp. nov.

The above research would require improved sampling of the species' distributions, underpinned by verifiable biological (herbarium) collections. Combined with the need for increased natural history, ecological and genetic observations we would propose that the genus be the focus of a dedicated field campaign.

Victoria also has great potential to serve as a valuable model for exploring the biogeography of continental South America as it is an aquatic species mostly restricted to large river systems; its seeds are desiccation intolerant and thus unable to escape flooding plains of their water catchment. The ephemeral nature of such flooding plains (Cowgill and Prance, 1989) is thought to have driven gigantism, as a mode of outcompeting other aquatic plants (Box et al., 2022). *Victoria* could feasibly be part of a monophyletic clade comprising *Microvictoria-Euryale-Victoria,* as is suggested by morphological observations (Gandolfo et al., 2004), the age of which predates the complete break up of Gondwana ca 83 Ma (Seton et al., 2012). The current distribution of genus *Victoria* spans a vast area of river systems, representing ca 44% of the South American drainage basins (ca 7.8×10^6 km^2) in a region where orogenesis and changing climates have wrought major changes in the last 12 Ma (Antonelli et al., 2009; Figueiredo et al., 2009; Hoorn et al., 2010, 2022).

TAXONOMY

Key to the Species

1. Mature leaves with upturned rim, rim moderate to high (8–10% of blade length), sigmoid in cross- section; mature bud concave towards apex; carpellary appendages with a cuneate base, arising 45° from the point of attachment (see **Figure 2**); prickles associated with the flowers tapering abruptly to a point, covering ovary and either absent, or covering the basal 1/3 of abaxial surface of the outer tepal, prickles 0 – 100 per tepal; seeds globose, raphe faintly visible. Paraná river basin, lower course of Paraguay river *V. cruziana* (**Figure 6**)

1. Mature leaves with no upturned rim, or where present the upturned rim low to moderate (4–7% of blade length), sigmoid or vertical in cross-section; mature bud convex towards apex; carpellary appendages with an auriculate or subauriculate base, arising 45° from point of attachment or not (see **Figure 3**); prickles associated with the flowers tapering smoothly or abruptly to a point, covering both the ovary and entire abaxial surface of the outer tepal, or only the ovary, 0–1000 prickles per tepal; seeds ellipsoid or globose, raphe faintly visible or prominent. Amazon river basin, Pantanal 2

2. Mature leaves with no or moderate upturned rim, which, where present, is vertical in cross-section. Prickles associated with the flowers tapering smoothly to a point and covering both ovary and entire abaxial surface of the outer tepals, 55–300 prickles per tepal; stigmatic chamber deeply concave, obdeltate in longitudinal profile, carpellary appendage auriculate and hanging free from, not arising 45° from point of attachment (see **Figures 2**, **3**); seeds ellipsoid, the raphe faintly visible. Amazon river basin excluding *V. amazonica* (**Figure 4**)

2. Mature leaves with a moderate upturned rim, which is sigmoid in cross-section. Prickles associated with the flowers tapering abruptly to a point, covering both the ovary and entire abaxial surface of the outer tepals, or only the ovary, 0–1000 prickles

per tepal; stigmatic chamber shallowly concave and oblong in longitudinal profile, carpellary appendage subauriculate at point of attachment not arising 45° from point of attachment (see **Figure 3**) (not known for *V. cruziana* f. *mattogrossensis*); seed ovoid, the raphe prominent. Llanos de Moxos or Pantanal 3

3. Prickles associated with the flowers covering the ovary and either absent from or sparsely distributed over the abaxial surface of the outer tepals, 0–10 per tepal; apical portion of the carpellary appendage longer than the basal portion. Llanos de Moxos *V. boliviana* (**Figure 5**)
3. Prickles associated with the flowers covering the ovary and distributed densely and evenly over the abaxial surface of the outer tepals, 500–1000+ per tepal; apical portion of the carpellary appendage shorter than the basal portion. Pantanal *V. cruziana* f. *mattogrossensis taxon incertum taxon incertae*

Victoria R. H. Schomb., *Athenaeum* (London) 1837 (No. 515): 661 (September 9 1837).

Victoria Lindl., Monog. 3 (October 16 1837).

Victoria J. E. Gray, Mag. Zool. Bot. 2(10): 373 (December 1 1837).

Aquatic perennial herb, rhizome erect, tuberous, elongate to cylindrical, roots adventitious. Leaves floating, orbicular, peltate, perforated by stomatodes, adaxial surface of lamina glabrous, lacking prickles, green; abaxial surface of lamina with prominent radial and reticulate ribs, juvenile leaves sagittate; leaf margins flat or upturned; prickles covering petiole and ribs. Inflorescences uniflorate, bracteate. Flowers axillary, solitary, multiple buds per plant; pedicel with 4 primary air chambers, 8 minor chambers, covered in prickles; flowers opening one at a time, projecting above water surface shortly before anthesis, projecting above or resting on water surface at anthesis, each flower opening over two nights and partially closing in between, protogynous. Epigynous, ovary globose, covered in prickles externally, ovules parietal, attached by short funiculi, globose. Outer tepals 4, triangular, apex acute to rounded. Inner tepals 40–c.100, arranged in spiral series, creating a torus (attachment point of tepals forming a ring of tissue), tepals gradually reducing in size towards the center and changing shape from apically rounded to acute from outer to innermost; outer staminodia in 1 or 2 whorls, thick, rigid, apiculate; stamens > 100, borne in c. 3 series, subulate, introrse; anthers linear-elongate; inner staminodia, >50, sigmoid, subulate, partially adnate to carpellary appendages, detaching at second-night anthesis; carpellary appendages L-shaped, arising from extension of stigmatic surface, lower parts adnate to tissue extending from tepal base attachment, corresponding in position and number with stigmatic surface ridges and locules. Fruit ripening just below surface of water, 10–15 cm in diameter (excluding prickles) at maturity, fleshy, oblate, topped by a shallow cylinder-shaped mass of dark reddish to maroon ring of persistent hard tissue formed by the remnant bases of tepals; inner staminodia persistent and curved over concave stigmatic surface while ripening; outer layers of pericarp disintegrating to release seeds. Seeds smooth, surrounded by a mucilaginous aril.

Three, possibly four species, tropical and temperate South America.

Victoria amazonica (Poepp.) Klotzsch, *Bot. Zeitung (Berlin)* 5: 245 (1847). *Euryale amazonica* Poepp. *Froriep's Not. Natur-Heilk*. 35: 131 (1832). Type: *Poeppig s.n.* (holotype W -presumed destroyed in WWII); Brazil, Amazonas, Careiro da Várzea [Teresina], Ilha de Careiro, 25 Sept. 1974, *G.T. Prance* 22745 (neotype (selected by de Lima et al., 2021): INPA (INPA46745); isoneotypes: K (K000837777!), NY (NY2269910, NY2269911, NY2269928), MO (MO3414212), US (US01341606)). Vernacular names: Forno de Jaçanã, Auapé yapóna, *Victoria regia*, Giant Amazonian Waterlily. **Figures 1A**, **2**, **3A–C**, **4**, **9**.

Victoria regina R.H.Schomb. Athenaeum (London) 515: 661 (September 9, 1837).

Victoria regia Lindl., Monograph: 3 (October 16, 1837). Type: Victoria Regia: 3, Plate 1 (October 16, 1837). *nom. superfl.*

Victoria regina J. E. Gray, *Mag. Zool. and Bot*. 2(11): 440 (December 1, 1837). *nom. superfl. Victoria reginae Hook. Hooker's J. Bot. Kew Gard. Misc*. 2: 314 (1850). *orth. var.*

Leaves up to 2.3 m broad, adaxial surface of lamina green, occasionally tinged bronze in younger leaves; abaxial surface of lamina maroon or green, radial and reticulate ribs maroon, yellow or green; leaf margins form a low to moderate rim c. 4–7% of the lamina length (higher in crowded habitats), rim curved at its base then ± perpendicular to adaxial surface, abaxial surface of rim maroon or green; hairs 0.3–12 mm, simple, multicellular, 3–12 segmented. *Flowers* up to 28 cm in diameter at second-night anthesis. Ovary 8–12 cm diameter, outer surface covered in prickles 1– 18 mm (dried), prickles gradually tapering to a sharp point, hairs absent or present, where present simple, 0.1–0.4 mm, inner surface of ovary with deeply concave stigmatic surface, rounded to triangular in longitudinal profile, ridged with lines corresponding with 25–36 radially arranged locules, each containing 25–28 ovules, 1–1.5 mm diameter (fresh). Outer tepals 4.9–12 × 4–8 cm when fresh; abaxial surface predominantly brown/maroon, bearing 55–330 prickles per tepal, prickles tapering gradually to a sharp point, ranging from 1–14 mm (dried), spaced regularly, irregularly, or clustering more densely toward the base over entire surface, hairs absent or present on abaxial surface, where present 0.1–0.2 mm. Inner tepals 7–15 × 2–6 cm (fresh), innermost deep maroon in bud; all others remaining white or turning pink to dark pink at second-night anthesis; outer staminodia > 25, 5–6 × 1–1.5 cm thick, rigid, apiculate; stamens 2–4 × 0.5–1 cm; inner staminodia, 4–6 × 0.5–1 cm; base of lower parts of carpellary appendage auriculate/rounded in shape and hanging free from extension of stigmatic surface, length of upper parts not exceeding that of lower parts. *Flower at first night of anthesis,* inner tepals white, with innermost tepals dark maroon, outer staminodia tipped pink; *second night anthesis,* innermost tepals dark maroon, inner tepals remaining white or pink to dark pink or red, darkest at base, outer staminodia remaining white or dark pink for basal two thirds of their length, tipped pink, inner staminodia pink at base. *Seeds* 600–1000 per fruit, 7–8 × 9–10 mm, ellipsoid, green to brown, raphe faintly visible.

Distribution and Conservation Status —*Victoria amazonica* is restricted to the Amazon river basin, from Northern

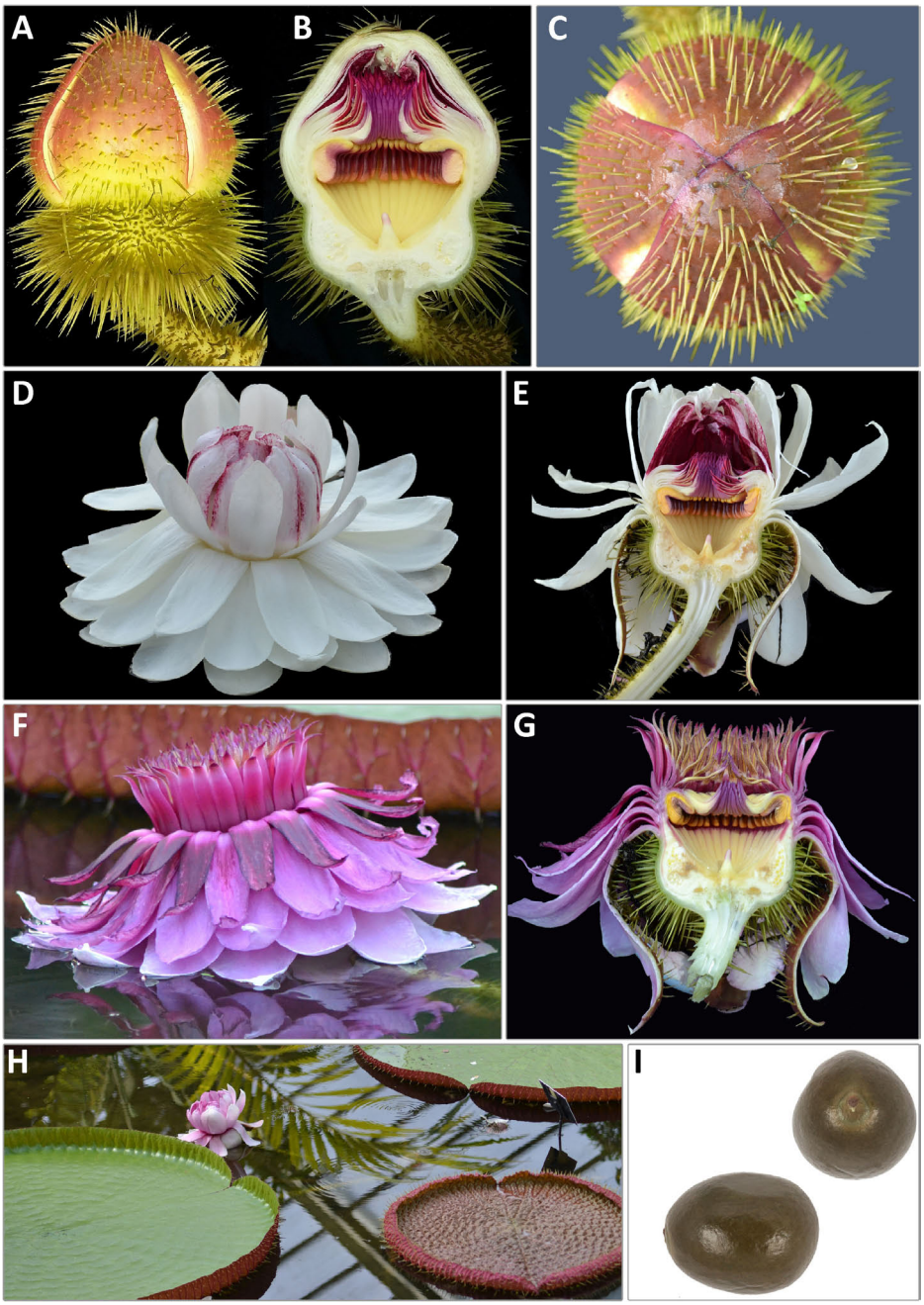

FIGURE 9 | *Victoria amazonica* **(A)** bud whole, **(B)** bud L.S., **(C)** bud from above, **(D)** first night flower, **(E)** first night flower L.S., **(F)** second night flower, **(G)** second night flower LTS, **(H)** habit, and **(I)** seed. **(A–H)** (LTS) and **(I)** (CM) cultivated RBG Kew.

Brazil, Bolivia, Colombia, Guyana and Peru. Its EOO is estimated to be 2,640,795 km^2, exceeding the threshold for an IUCN threat category under criterion B, whilst its AOO is estimated as 476 km^2, falling into the Endangered category. We believe that our calculation of the AOO is likely an underestimate, resulting from difficulties in observing the species in the field and under-representation of the genus in biological collections.

There are more than 10 locations for which threats have been assessed, but there is not the information on population fragmentation or fluctuation available in order to be able to assess the 'severely fragmented' and 'extreme fluctuations' subcriteria. A continuing decline in habitat quality is inferred due to the presence of hydroelectric dams, mining, and deforestation of the river systems from where *V. amazonica* is documented. For example, in Peru, localities along the Marañón river to

the headwaters of the Amazon and the Ucayali river have been heavily deforested by gold mining (i-Terra, 2021). Gold mining is associated with profound mercury contamination of rivers and aquatic species (USGS Environmental Health Program, 2019) and increases in Peruvian mercury imports suggest that contamination must be increasing (Swenson et al., 2011). In Brazil, as in Bolivia there have also been reports of mining activities in indigenous lands adjacent to several *Victoria* populations (Hutukara Associação Yanomami and Associação Wanasseduume Ye'kwana, 2020; INPE, 2021; MapBiomas, 2021; Mercado, 2021).

Victoria amazonica is here assessed as Least Concern (LC) considering that its range stretches across Amazonia and currently exceeds the parameters for a threatened category under criterion B. We note, however, a moderate number of locations where populations are under threat and there is a continuing decline in habitat quality. Further investigation and surveys are needed to better understand trends in population size, fragmentation, distribution and the impact of climate change.

Notes. *Victoria amazonica* is the only species whose leaves do not always form an upturned rim, and when they do, it is usually low and vertical in profile rather than recurving over the flat part of the lamina. Its flowers are distinguished from *V. cruziana* and *V. boliviana* both in bud and on first-night opening, as the innermost tepals are dark maroon rather than white. Carpellary appendages are curved at the base of the lower part and hang freely away from the attachment point. The prickles covering both outer tepal abaxial surface and outer ovary are uniquely gradually tapering to a sharp point (not abruptly tapering as in other species), and always cover the entire abaxial surface of the outer tepals. Its seeds are ellipsoid rather than globose.

Material Examined— BOLIVIA. **Pando:** Manuripi: pond north of Rio Madre de Dios., –66.126667, –10.903611, 09/07/1997, *Ritter, N., Crow, G.* & *Crow, C. 4170* (LPB, MO). BRAZIL. 'North Brasil':

1898, *Vaughan, G. 61* (K). **Acre:** Río Moa, margem esquerda; lugar chamado Humaita, –72.895556, – 7.614444, 01/10/1984, *Ferreira, C. A. 5123* (NY). **Amapá**: Itaituba, Igarape no Rio Tapajos, – 55.961111, –4.229167, 16/12/2017, *Brogim, R. 4* (UPCB). **Amazonas:** Ipixuna, Margem do Rio Croa, – 72.556667, –7.745278, 15/02/2009, *Quinet, A., Saraiva, B., Firmeza, T.1582* (K, SPF); Teresina, Ilha de Careiro, –59.81667, – 3.1, 25/09/1974, *Prance, G. T. 22745* (K, NY, US); Basin of Rio Purus area. Lago Preto, 3 km north of Labrea, – 64.813056, –7.229772, 29/10/1968, *Prance, G. T., Ramos, J. F. and Farias, L. G. 8016* (NY); Rio Solimões, south bank near Carreiro, –59.806667, –3.168889, 05/02/1974, *Steward, W. C. And Ramos, J. F. P20211* (K, NY, US); Rio Amazonas, from Manaus to 100 km lower reaches, –59.141944, –3.216111, 08/08/1987, *Tsugaru, S. and Yotaro Sano B-769* (MO, NY); Ilha do Cantagalo, –61.503889, –1.570833, 04/07/1995, *Adalardo-Oliveira, A. 2645* (NY, SPF); Riverside and small islets of Rio Solimões within 100 km upper-stream from Manaus, – 60.728056, 3.260278, 15/08/1987, *Tsugaru, S. and Yotaro Sano B-1069* (NY); **Pará**: Santarem, Igarape, Ilha Grande de Santarum, –54.706840, –2.450070, –/10/1849 and –/11/1849, *Spruce, R. 441* and *Spruce s.n., s.d.,* (K, M, P), – /04/1850(NY) 1849 (P); Oriximina, Lago Uraria, SW of Orixima, across Rio Trombetas, – 55.9025, –1.812778, 11/06/1980, *Davidson, C. and Martinelli, G. 10241* (MO, NY, RB, US); Rio Cupari, Lago Curuca, 01/01/1948, *Black, G. A. 48-2223* (IAN); Rio Cupari, Lago Curuca, 02/01/1948, *Black, G. A. 48-2253* (IAN); Rio Cupari, Lago de Curuca, 02/01/1948, *Black, G. A. 48-2254* (IAN, NY, US); Oriximina, Rio Trombetas, Lago Ururia, 6 km SW de Oriximina, –55.910278, –1.803889, 08/06/1980, *Martinelli, G. 6945* (RB); Pacoval, Rio Curua, –55.083333, –1.833333, 6–8/08/1981, *Jangoux, J., and Riberio, B. G. S. 1647* (NY); Santarem, 25/12/1938, *Markgraf 3873* (RB); Monte Alegre, Rio Gubatuba, proximo a vila Pare Sol, –54.043333, –2.011944, 17/07/2011, *Lima, C. T. 503* (HUEFS); Pacoval, Rio Curua, –55.083333, – 1.833333, 6–8/08/1981. **Roraima:** Rorainopolis, Rio Branco, Lago do Pirarucu, 25 km antes da boca com o Rio Negro, – 61.8525, –1.158333, 28/03/2012, *Martinelli, G., Moraes, M. A., Benevides, P., Forzza, R. C., Nadruz, M., Gallucci, S., Costa, D. 17700* (RB). COLOMBIA. **Amazonas**:Leticia, below Quebrada de Arara, –70.065278, –4.05944, 28/01 – 07/02/1969, *Plowman, T., Lockwood, T., Kennedy, H., Schultes, R. E. 2313* (K). GUYANA. **Berbice:** Berbice, –58.2778, 4.394033, 1837, *Schomburgk s.n.* (K). **Upper Takutu-Upper Essequibo**: Karanambo, Rupununi River, –59.3, –3.75, 27/09/1988, *Maas, P. J. M., Koek-N, J., Lall, H., ter Welle, B. J. H., Westra, L. Y. 7727* (K). PERU. **Maynas:** East of Puerto Alegria, –70.0625, –4.103611, 15/03/1977, *Gentry, A. and Daly, D. 18351* (MO); Isla Padre (Cocha Paster), –76.166667, – 3.75, 21/12/1982, *Vasquez, R.., Grandez, C. and N. Jaramillo, N. 3684* (MO); Padre Isla in Río Amazonas, and in the cato below Iquitos., –73.163611, –3.651944, 22/05/1978, *Gentry, A., Jarmillo N. 22133* (MO). See **Supplementary Data**.

Victoria boliviana Magdalena and L. T. Sm., ***sp. nov***. Type: Bolivia, Beni Department, Provincia Ballivían, subiendo el Río Yacuma desde Puerto Espíritu, laguna en conexíon al Río Yacuma, unos 20 m al N, 29 Mar. 1988, *S. G. Beck 15173* (holotype: LPB; isotype: K (K000798309). Vernacular names: Reina Victoria, Victoria regia. **Figures 1B**, **3D–F**, **5**, **10**.

Most similar to *V. cruziana* Orb., from which it can be distinguished by the lower rim of the floating leaf, convex apex of the flower bud, length of the upper part of the carpellary appendages exceeding that of the lower part and the larger seeds. The *V. boliviana* plastid genome differs from that of other *Victoria* species by a 14 bp insertion between plastid genes *ndhC* and *trnV* in the large single copy region (LSC), a 5 bp deletion between *trnK* and *rps16*, a 7 bp deletion adjacent to *trnC* in the LSC and a 42b p deletion in the CDS of gene *ycf1*, within the SSC. Finally, a 4 bp transversion unique to *V. boliviana* sp. nov. was found in the LSC.

Leaves up to 3.2 m broad, adaxial surface of lamina green; abaxial surface of lamina dark green, maroon or dark-blue, radial and reticulate ribs yellow or green; leaf margins upturned to form a moderate rim c. 4–7% of leaf length, rim recurving strongly over blade surface at base and curving inwards or flared outwards at top, abaxial surface of rim deep maroon or very pale green/white in color, glabrous or with hairs, where present 1.2–3 mm, simple, multicellular, 6–15 segmented. *Flower* bud broadly ovoid, convex at apex, up to 36 cm in diameter at second-night anthesis. Ovary

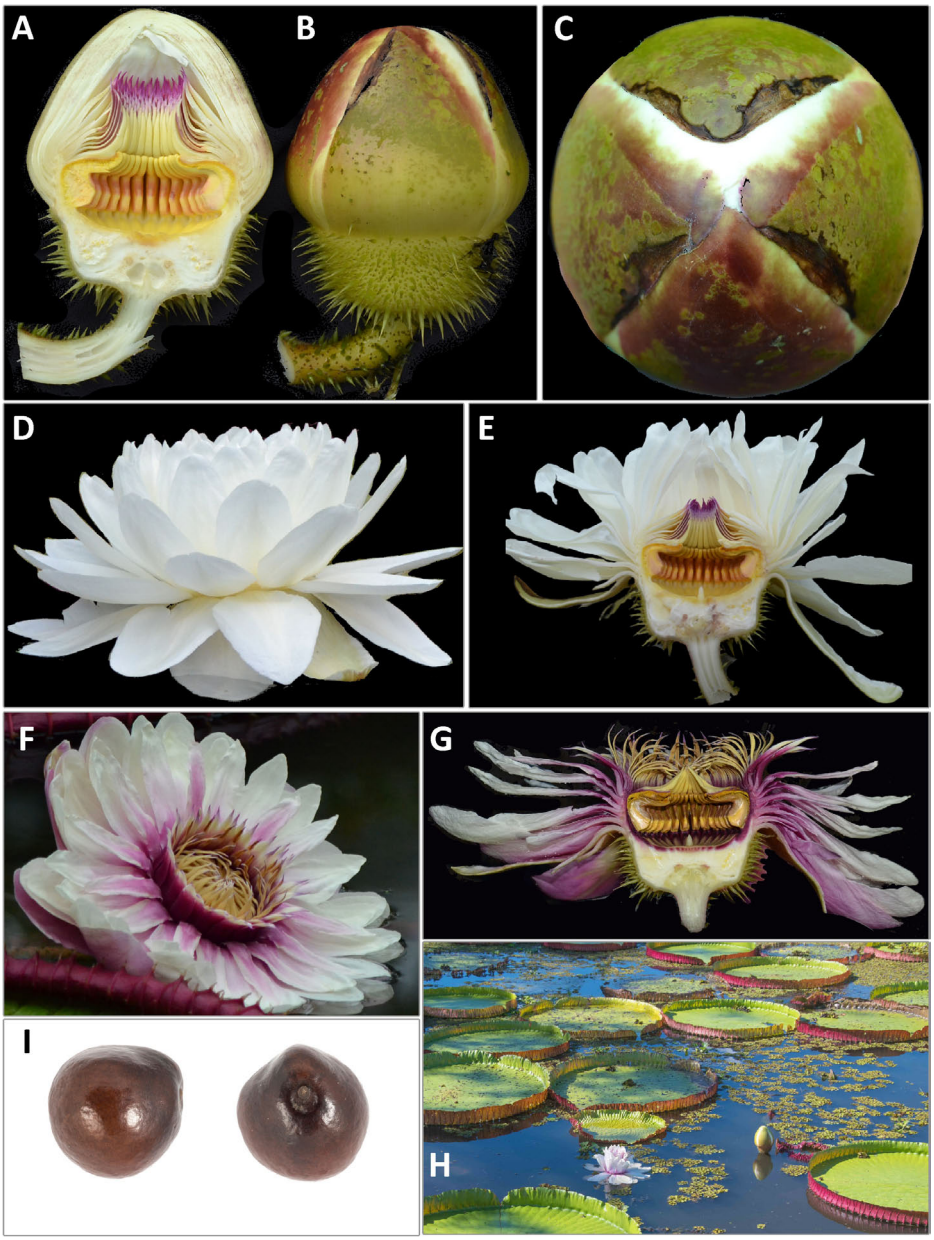

FIGURE 10 | *Victoria boliviana* sp. nov. **(A)** bud whole, **(B)** bud L. S., **(C)** bud from above, **(D)** first night flower, **(E)** first night flower L. S., **(F)** second night flower, **(G)** second night flower L.S., **(H)** habit, and **(I)** seed. **(A–G)** (LTS), **(I)** (CM) cultivated RBG Kew, H (CM) Beni, Bolivia.

8–10 cm diameter, outer surface covered in prickles, 1–10 mm (dried) glabrous; prickles abruptly tapering from c. half of length to sharp apex; inner surface of ovary with shallowly concave stigmatic surface, oblong in longitudinal profile, ridged with lines corresponding with 25–36 radially arranged locules, each containing 8–14 ovules, 2–2.5 mm diameter (fresh). Outer tepals 4, 10–15 × 8–10 cm when fresh, abaxial surface predominantly green, or tinged maroon, prickles absent or present, where present up to 10 per tepal, prickles tapering abruptly at their midpoint to a sharp point, 1–10 (dried), distributed irregularly over entire surface, glabrous. Inner tepals 6–15 × 1.5–9 mm (fresh), innermost remaining white or turning pale pink at their base at the second-night anthesis; outer staminodia > 50,3–4 × 0.5– cm, thick, rigid, apiculate; stamens, 4–5 × 0.5–1 cm; inner staminodia 4– 5 × 0.5–0.7 base of lower parts of carpellary appendage angular in shape and arising at 45 degree angle from stigmatic surface, length of upper parts exceeding that of lower parts. *Flower at first night of anthesis,* inner tepals white, outer staminodia tipped blue-violet; *second night anthesis*, inner tepal adaxial surface pink, inner tepals pale pink at base, white or pink towards the apex, outer staminodia dark pink for basal two-thirds of their length, white then violet towards the apex, inner

FIGURE 11 | *Victoria cruziana* **(A)** bud whole, **(B)** bud L.S., **(C)** bud from above, **(D)** first night flower, **(E)** first night flower L.S., **(F)** second night flower, **(G)** second night flower LTS, **(H)** habit, and **(I)** seed. **(A–D,F,G)** (photo LTS), **(I)** (CM) cultivated RBG Kew, **(D,E)** cultivated Denver Botanic Gardens (LTS).

staminodia pink at base. *Seeds* c. 300 per fruit, 12–13 × 16–17 mm, globose with a prominent raphe (especially when dry), dark brown to black, surrounded by a mucilaginous aril.

Distribution and Conservation status *Victoria boliviana* Magdalena and L. T. Sm. is restricted to Bolivia and the flood plains of the Llanos de Moxos, Mamoré watershed, identified as a Centre of Plant Diversity and Endemism (Site SA24) (Beck and Moraes, 1997). These area is considered by Langstroth Plotkin (2012). Moxos is surrounded by the forests of the Upper Madeira basin and is an area of largely open vegetation – herbaceous wetlands, grasslands, savannas, and woodlands (Beck, 1983, 1984; Langstroth Plotkin, 2012). Images (not included in the minimum calculation of the EOO or AOO) suggest that *V. boliviana*'s range extends further west (natural or cultivated) to Rurrenabaque. We estimated a minimum and maximum EOO and AOO. The maximum range was based on the

potential habitat across the Llanos de Moxos region (including geographical information from public unverified images) and verified herbarium collections and iNaturalist images. The minimum range was based on the coordinates of herbarium collections and iNaturalist images alone. We estimated that the EOO of *V. boliviana* ranges between 8,006 km^2(minimum) and 33,151 km^2 (maximum), falling close to the thresholds of the Vulnerable category under criterion B. We estimated that the AOO of *V. boliviana* ranges from 32 km^2 (minimum) to 2,000 km^2 (maximum), falling between the Endangered and Vulnerable categories. There are less than five known locations for *V. boliviana*. We could find no information about population fragmentation but believe that *V. boliviana* may be vulnerable to fluctuations in flooding and drought throughout the year. For example, Beni Department has been recently affected by seasonal floods, fires and droughts due to El Niño and La Niña climate events (Vásquez, 2015). A recent increase in agriculture-lead deforestation has been documented along the Trinidad-Santa Cruz highway (Langstroth Plotkin, 2012), to the south of known *V. boliviana* populations and satellite images from Google Earth Pro and i-Terra suggest extensive deforestation along the edges of roads (i-Terra, 2021; NASA, 2021; GoogleEarth, 2022) which we use to infer an active decline in habitat quality.

Taking a precautionary approach and based on the small EOO and AOO, small number of locations (5), and continuing decline in habitat, we assess *V. boliviana* as Vulnerable (**VU**), according to criteria B1ab(iii)+B2ab(iii).

Notes. *Victoria boliviana* sp. nov. has the largest observed leaves of the three species, with laminae > 3 m in length having been observed. The abaxial surface of the upturned rim surface varies between individual plants in the same locality from dark maroon to very pale, almost "white" green, a characteristic not found in other species. Prickles on abaxial outer tepal surfaces are absent or very few, and if present are not confined to the lower portion of the outer tepals unlike *V. cruziana* where the smaller number of prickles are confined to the lower one-third of the abaxial tepal surface.

Victoria boliviana is the only species of Victoria whose carpellary appendages have upper portions that are longer than the lower portions. In addition, *V. boliviana*'s stigmatic chambers are the shallowest of the three *Victoria* species. Haenke saw *Victoria* plants in Yacuma in 1801, during the Malaspina expedition (Gickelhorn, 1966, p. 105; Ibañez Montoya, 1984), but no identifiable description is available and no voucher has been found. d'Orbigny (1840, p. 57) reported seeing this species on the banks of the Mamoré river in 1832 and mistakenly assigned this species to *V. amazonica* when publishing his description of *V. cruziana*.

Further investigations and surveys are required to better understand the species' current range, population fluctuations and habitat and thus better predict the impacts of the threats identified.

Field observations of the flowers from a single population in the Llanos de Moxos suggest that whilst pollinated by beetles, *V. boliviana* sp. nov. flowers may host fewer individuals of pollinators than *V. amazonica*, only 4–10 individuals being observed in the flowers of the former, compared to > 20 in the flowers of the latter. This could be due to a lower density of pollinators in their area of occupation.

Additional Material – BOLIVIA. **Beni:** Santa Ana del Yacuma, –65.4236, –13.74, –/6–7/1845, *Bridges, T. s. n.* (K); Cercado: Laguna Suarez 5 km sur de la ciudad de Trinidad, –64.864167, – 14.872222, 08/05/2019, *Magdalena, C. Melgar, D. G., Salazar, C. D., Alvarez, C., Gutierrez, G., Arias, J. 154* (German Coimbra Sanz Herbarium, Jardin Botanico Municipal); Moxos, pasando el Río Mamoré, cerca al puente del Río Tijamuchi a lado del camino, –65.145278, –14.851111, 09/05/2019, *Magdalena, C. Melgar, D. G., Salazar, C. D., Alvarez, C., Gutierrez, G., Arias, J.155* (Herbario German Coimbra Sanz Jardín Botanico Municipal).

Victoria cruziana Orb., *Ann. Sci. Nat., Bot., sér. 2*, 13: 57 (January 1840). Type: Bolivia [Argentina], Corrientes, banks of the Paraná river, Arroyo de San José, beginning of 1827, d'Orbigny *s.n.* (lectotype: P (P02048598*) (designated by de Lima et al., 2021); isolectotypes: P (P02048599).

Vernacular names: Irupé (yrupé), yacare yrupé, naanók lapotó (poncho del otany), maíz de agua, Santa Cruz Waterlily, Victoria regia. **Figures 1C**, **3G–I**, **6**, **11**.

Victoria regia var. *cruziana* (Orb.) G. Lawson, *Proc. & Trans. Roy. Soc. Canada* 6(4): 109 (1889)

Euryale brasiliana Steud., *Nomencl. Bot.* [Steudel], ed. 2. 1: 617 (November 1840).

Euryale policantha Rojas Acosta, *Cat. Hist. Nat. Corrientes* 65 (1897).

Euryale bonplandia Rojas Acosta, *Cat. Hist. Nat. Corrientes* 151 (1897).

Victoria cruziana f. *trickeri* Henkel ex Malme, *Acta Horti Berg.* 4(5): 12. 1907 ("Trickeri"), *nom. nud.*

Leaves up 2.4 m broad, adaxial surface of lamina green, abaxial surface of lamina green or dark blue-green, radial and reticulate ribs yellow or green; leaf margins form a high rim 8–10% of leaf length, rim ± perpendicular to or slightly recurved over adaxial surface at base, flared outwards at top (sigmoid in profile), abaxial surface of rim green or tinged maroon, hairs 1–3 mm, simple, multicellular, 10–15 segmented. *Flower* bud broadly ovoid, concave just before apex, up to 30 cm diameter at second-night anthesis. Ovary 7–10 cm diameter, outer surface covered in prickles 1–22 mm (dried), prickles abruptly tapering from c. half their length to sharp apex; hairs absent or present, where present 0.1–12 mm; inner surface of ovary with moderately concave stigmatic surface, rounded to triangular in longitudinal profile, ridged with lines corresponding with 25–38 radially arranged locules, each containing 20–25 ovules 1.5–1.8 mm (fresh). Outer tepals 4, 10–13 × 4–9 cm when fresh, abaxial surface green and/or tinged maroon abaxially prickles absent or present, where present up to 100 per tepal, prickles tapering abruptly at their midpoint to a sharp point, 1–10 mm (dried), distributed up to lower one third of surface, hairs absent or present, where present 0.1–1 mm. Inner tepals 7–10 × 1.5–9 cm (fresh), innermost all white both in bud and during first-night anthesis, crinkled in appearance, turning pale to dark pink on second-night; outer staminodia, 6–7 × 1–1.5 cm, thick, rigid, apiculate; stamens 4–6 × 0.5–1 cm

inner staminodia > 50, 4–6 × 0.5 cm; base of lower parts of carpellary appendage flat, arising from stigmatic surface at 45 degree angle, cuneate, length of upper parts not exceeding that of lower parts. *Flower at first night of anthesis:* all inner tepals white, outer staminodia tipped pink; at *second night anthesis,* outer tepal adaxial surface pink, inner tepals pale or dark pink at base, white or pink towards apex, outer staminodia dark pink for basal two-thirds of their length, white then pink towards apex, inner staminodia pink at base. Seeds, c. up to 1000 per fruit, 7–9 × 8–10 mm, globose, raphe faintly visible, brown to black, surrounded by a mucilaginous aril.

Distribution and Conservation Status

Victoria cruziana f. *mattogrossensis taxon incertum* has hitherto been included within *Victoria cruziana* until this study. Because our genomic analyses are limited with respect to the rank of this taxon and of its relationship to the other species, we have not considered forma *mattogrossensis taxon incertum* as conspecific with *V. cruziana* for the purposes of an extinction risk assessment.

Victoria cruziana is restricted to the Paraná river basin and tributaries, from Paraguay to Argentina, and possibly Bolivia. Based on the maximum potential habitat of wetlands (including Esteros de Ibera National Park Wetlands) and a combination of verified herbarium collections and iNaturalist images we estimate the EOO of *V. cruziana* to be between 46,563 and 132,945 km^2. This exceeds the threshold for a threatened category under criterion B (IUCN Standards and Petitions Committee, 2019). We calculate the AOO to be 120 km^2, although an upper estimate based on the extent of the river may exceed 2,000 km^2 (but not more than 3,000 km^2). This would assess *V. cruziana* as between Endangered and Vulnerable categories. There are more than 10 locations but no information about population fragmentation. We infer a continuing decline in habitat quality due to the increasing frequency of droughts, the abstraction of water, deforestation (Caivano and Calatrava, 2021; Comisión Nacional de Actividades Espaciales, 2021) and big hydroelectric dams. For example, Itaipu is one of the largest dams in the world (Stevaux et al., 2009) and lies within the *V. cruziana* range. Whilst the AOO upper estimate is close to the threshold for VU, there are more than sufficient criteria for a threatened category under criterion B, and we therefore assess *V. cruziana* as LC.

We recommend further documentation of the distribution and size of *V cruziana* populations and investigations into their fluctuation through time as we believe that it may be vulnerable to an increase in the frequency and severity of droughts associated with climate change and increased sedimentation caused by the construction of large dams within its habitat (Stevaux et al., 2009).

Notes. *Victoria cruziana* forms the proportionately highest leaf rims of all the species, and these are always slightly recurved over the flat part of the lamina, flaring out at the top. The concavity of the outer tepals before their apex gives the bud a pinched-in appearance. Prickles are absent or occur on the outer tepal abaxial surface, but only up to one-third of their length from the base. In this species, hairs which are sometimes present on the lower outer tepal abaxial surface and ovary are the only ones large enough to see without magnification. At first night anthesis, inner tepals have a crinkled appearance.

Victoria cruziana f. *mattogrossensis taxon incertum* was described by Malme based on material present in the spirit collection of the Swedish Museum of Natural History comprising two jars of the same gathering. S07-84 comprises a longitudinal section through the ovary and perianth showing clearly the distribution of prickles and their form on both the ovary and outer tepals, also the morphology of the carpellary appendages; S07-85 comprises sections of the petiole apex and either the petiole or pedicel.

Both are preserved in excellent condition. S07-84 was selected as lectotype as it displayed a greater number of diagnostic morphological features.

Victoria cruziana f. *mattogrossensis taxon incertum* was described from, and material corresponding to it has only been observed from the Pantanal, in Bolivia, Brazil and Paraguay in the Uruguay river basin.

Phylogenomic data were unable to confirm whether the material sampled represents a distinct evolutionary lineage (see section "Discussion"). It may be that further sampling and research supports the recognition of this name as a distinct taxon.

Material Examined. ARGENTINA. **Chaco:** 1st De Mayo, Laguna en Chacra, al lado del arroyo Ine, –58.849444, –27.416389, 02/03/2006, *Mulgura de Romero, M. E., Anisko, T., Harbage, J., Illarrage, H.4249* (SI); San Fernando, Entre Barranqueras e Isla Antequera, –58.87944, –27.416389, 18/03/1967, *Krapovickas, A. and Cristobal, C. L. 12752* (MO). **Corrientes:** Capital, Riachuelo, off Ruta 12 ca 17 km S of Corrientes, –58.749444, –27.554444, 06/04/1982, *Schinini, A. Wiersema, J. H. 2243* (MO); Esquina, Isla Correntina frente a curuzu – Chali, en el Paraná medio, –59.630833, –30.341944, 10/04/1968, *Burkart, A., Troncoso, N. S., Guaglianone, E. R. and Palacios, R. A. 26963* (SI); San Roque, R. Santa Lucia, –58.738056, –28.576944, 27/02/1957, *Pedersen, T. M. 4486* (K, MO); Bella Vista, Cruce Ruta Nacional 12 Ey Puente Sobre Lel Río Santa Lucia, –58.72, –28.57, 12/04/2008, *Mulgura de Romero, M. E., Aniśko, T., Belgrana, M. J. and Harbage, J. 4474* (SI); Dep. Esquina, Río Guayguiraro, –59.56, –30.374444, 26/02/1974, *Quarin, C., Schinini, A. Gonzales, J. M., Ishikawa, A. 2196* (K). **Santa Fe:** La Capital, Ruta Nacional 168 y Puente no. 9, Sobre Arroyo mini, al E-SE de la Ruta Prov. 1. W, –60.575833, –31.672778, 12/04/2008, *Mulgura de Romero, M. E., Anisko, T., Belgrana, M. J. and Harbage, J. 4477* (SI). BOLIVIA. **Santa Cruz:** Angel Sandoval, channel at southern end of Laguna Mandiore, –57.483333, –18.216667, 16/07/1998, *Ritter, N., Crow, G. E., Garvizu, M. and Crow, C. 4562* (LPB, MO) [f. mattogrossensis], *Ritter, N., Crow, G. E., Garvizu, M. and Crow. C.* 4560 (LPB, MO). BRAZIL. **Mato Grosso du Sul:** Corumba, Cacimba da Saude, – 57.664167, –18.99833, 13/12/2002, *Avellar, A. L. F. 13* (COR) [f. Mattogrossensis]; Ladario, terminal da Branave, no porto de Ladario, –57.584722, –19.0225, 12/08/1994, *Sanches, A. L., Bortolotto, J. M., Damascenos Jr., G. 44* (COR) [f. mattogrossensis]; Corumba, Ladario, CODRASA Brejo (swamp), – 57.516389, –19.021111, 27/11/2004, *Souza Jr., A. F.,*

and Siqueira, C. S. 39 (COR) [f. Mattogrossensis]; Corumba, Ladario, –57.580556, –19.001111, –/07/1894, *Anon s.n. [G. O. A. Malme] S07-785, S07-784*

(S). PARAGUAY. **Distrito Capital:** L'Assumption [in the swamps], –57.604167, –25.256667, 17/03/1875, *Balansa, R. 523* (K, P). **Central:** Piquete Cue, –57.666667, –25.166667, 29/11/2000, *Zardini, E. M. and Guerrero, L. 55181* (MO); Bay of Ascuncion, wetlands, –57.61922, –25.273333, 15/03/1988, *Ericsson, K. 577* (MO). **Neembucu:** Yataity, –58.040556, –26.788611, –/03/1975, *Walter, M. A. 86* (K); Pilar Garden Club S, –58.276389, –26.867778, 28/01/2005, *De Egea Juvinel, J., Pena-Chocarro, M.., Vera, M., Torres, M.. and Elsam, R. 738* (MO). **Presidente Hayes:** Riacho Pucu., –57.083333, –24.666667, 20/08/2000, *Zardini, E. M and Guerrero, L. 54765* (MO); 12/02/1987, *Sparre 2363/51* (P).

Taxon Incertum

Victoria cruziana f. *mattogrossensis* Malme, *Acta Horti Berg.* 4(5): 12. 1907. Type: [Brazil, Matto Grosso do Sul, near Corumba] *Anon s.n.*[G.O.A. Malme] 1894 [July] [S (spirit collection) (lectotype (selected here): S (S07-784*); isolectotype: S (S07-785*)]

Excluded Names

Victoria argentina Burmeister, *Reise durch die La Plata Staaten 2: 5* (1861).

Victoria fitzroyana Hort. Ex Loudon, Encyc. Pl. Suppl. Ii. 1388. = *Nymphaea gigantea* Hook.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: NCBI GenBank – SUB11368210.

AUTHOR CONTRIBUTIONS

CM, DGM-G, CDS, and GG-S conducted exploratory fieldwork and set up local partnerships in Bolivia which led to the conception of this study. LTS, CM, AKM, OAP-E, NASP, and AA conceived and designed the research, with contributions from OM and SD. CM, DGM-G, CDS, and GG-S conducted fieldwork. AKM conducted taxonomic and nomenclatural research. LTS created the botanical illustrations. LTS, CM, and AKM collected and analysed the morphological data. CM and LTS collected geographical data. NASP conducted wet lab work for DNA sequencing. OAP-E and NASP conducted *in-silico* molecular analyses. RN conducted conservation status assessments. IJL and SM conducted genome size wet lab work and analyses. SM conducted chromosome counts. CM cultivated living material that was used for horticultural observations, illustrations, the chromosome count and the estimation of genome size. SB, DGM-G, GR-G, and CDS contributed type material and plant tissue. AA contributed analytical tools and reagents. LTS and OAP-E produced the figures, with contributions from NASP. AKM, LTS, and NASP wrote the manuscript with contributions from CM, RN, OAP-E, and IJL. All co-authors reviewed and approved the submitted manuscript.

FUNDING

AA acknowledges financial support from the Swedish Research Council (2019-05191), the Swedish Foundation for Strategic Research (FFL15-0196), and the Royal Botanic Gardens, Kew.

ACKNOWLEDGMENTS

We thank Kit Knotts, curator of the Victoria-adventure website and collator of much of the gray literature on *Victoria* and whose website is the best source of trustable information on the Nymphaeaceae. She has helped to source plants for living collections, providing *Victoria* seeds of known pedigree to most of the botanic gardens in the world and has acted as a mentor to CM. We thank John Wiersema and Barre Hellquist for reviewing the taxonomy and ensuring the reliability of the Victoria-adventure website and the Paris herbarium for providing tissue samples of d'Orbigny's collection of *Victoria cruziana* for molecular study. We are grateful to Don Opitz for pointing us to Hooker's declaration that neither he or Lindley had seen Poeppig's publication prior to the publication of *Victoria regia*, Tomasz Aniśko for his magnificent book on *Victoria* and research into the genus's collection, description and subsequent cultivation, Sharon Willoughby (K) for comments on the historical context of *Victoria* in relation to RBG Kew, Johannes Lundberg (S) for providing us with high quality images of type material of *V. cruziana* f. *mattogrossensis*, Nicholas Hind (K) for advice on the typification of the species, Henke Beentje for advice on morphological terminology and the species' descriptions, Anders Lindstrom (Nongnooch Gardens) for the translation of Malme's revision of *Victoria cruziana*, and Anton Hagl for the translation of Poeppig's article in which *Euryale amazonic*a was described. Aaron Davis (K) and John Dransfield (K) for advice on morphotype description. We extend our gratitude to Eli Biondi, David Cooke, Tom Freeth, Will Spoelstra, Scott Taylor, Jean-Michel Touche, Alberto Trinco, and Mark Wilkinson from the Princess of Wales Conservatory (K), Solène Dequiret from the Waterlily House (K), and Tom Pickering and Paul Rees from the Tropical Nursery (K), for access to living material. We would like to thank Justin Moat (K) for help in producing the distribution maps and Steve Bachman (K) for reviewing and commenting on the extinction risk assessments; Thierry Deroin and Hubert Sinivassin (P) for providing images of the seeds of a d'Orbigny collection, as well as William Milliken (K), Mark Nesbitt (K), and Ghillean Prance (K) for help in sourcing ethnobotanical literature. We would also like to thank Matt Coulter, Stephen Kingdon, and John Sandham of the Botanic Gardens of South Australia, for providing tissue samples

for genomic study. A Dawn Jolliffe Bursary from the Royal Horticultural Society funded LTS's visit to the Denver Botanic Gardens where Tamara Kilbane and Mervi Hjelmroos-Koski from Denver Botanic Gardens provided access to living material for illustration.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.883151/full#supplementary-material

REFERENCES

Aniśko, T. (2014). *Victoria the Seductress*. La Jolla, CA: Beckon Books, 468.

Antonelli, A., Nylander, J. A., Persson, C., and Sanmartín, I. (2009). Tracing the impact of the Andean uplift on Neotropical plant evolution. *Proc. Natl. Acad. Sci. U.S.A.* 106, 9749–9754. doi: 10.1073/pnas.0811421106

Arbo, M. M., Lopez, G., Schinini, A., and Pieszko, G. (2002). "Las plantas hidrófilas," in *Flora del Iberá*, eds M. M. Arbo and S. G. Tressins (Corrientes: EUDENE), 9–10.

Archangeli, G. (1908). Studi sulli Victoria regia Lindl. Atti della Società toscana di scienze Botany, residente in Pisa. *Memorie* 24, 59–78.

Bachman, S., Moat, J., Hill, A. W., De La Torre, J., and Scott, B. (2011). Supporting red list threat assessments with GeoCAT: geospatial conservation assessment tool. *ZooKeys* 150:117. doi: 10.3897/zookeys.150.2109

Beck, S. G. (1983). *Vegetationsoekologische Grundlagen der Viehwirtschaft in den Ueberschwemmungs-Savannen des Rio Yacuma (departamento Beni, Bolivien)* Dissertationes Botanicae Bd. Vaduz: J. Cramer, 186.

Beck, S. G. (1984). Comunidades vegetales de las sabanas inundadizas en el NE de Bolivia. *Phytocoenologia* 12, 321–350. doi: 10.1127/phyto/12/1984/321

Beck, S. G., and Moraes, M. (1997). "Llanos de Mojos Region - Bolivia," in *"Centres of Plant Diversity. A Guide and Strategy for Their Conservation*, ed. S. D. Davis, V.H. Heywood and A.C. Hamilton (Cambridge: IUCN Publications Unit).

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

Borsch, T., Löhne, C., and Wiersema, J. (2008). Phylogeny and evolutionary patterns in Nymphaeales: integrating genes, genomes and morphology. *Taxon* 57, 1052E–1054E. doi: 10.1002/tax.574004

Borsch, T., Wiersema, J. H., Hellquist, C. B., Löhne, C., and Govers, K. (2014). Speciation in North American water lilies: evidence for the hybrid origin of the newly discovered Canadian endemic *Nymphaea loriana* sp. Nov.(Nymphaeaceae) in a past contact zone. *Botany* 92, 867–882. doi: 10.1139/cjb-2014-0060

Bortolotto, I. M., de Mello Amorozo, M. C., Neto, G. G., Oldeland, J., and Damasceno-Junior, G. A. (2015). Knowledge and use of wild edible plants in rural communities along Paraguay River, Pantanal, Brazil. *J. Ethnobiol. Ethnomed.* 11:46. doi: 10.1186/s13002-015-0026-2

Box, F., Erlich, A., Guan, J. H., and Thorogood, C. (2022). Gigantic floating leaves occupy a large surface area at an economical material cost. *Sci. Adv.* 8:eabg3790. doi: 10.1126/sciadv.abg3790

Byrne, D. (2008). In search of the dwarf *Victoria*. *Water Garden J.* 23, 11–14.

Caivano, V., and Calatrava, A. (2021). *Drought Hits South America River, Threatening Vast Ecosystem. Climate and Environment*. Toronto, ON: CTV News.

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

Comisión Nacional de Actividades Espaciales (2021). *Los Satélites Argentinos SAOCOM Monitorean la Bajante del Río Paraná. Ministerio de Ciencia, Tecnología e Innovación*. Buenos Aires, AC: Comisión Nacional de Actividades Espaciales.

Cowgill, U. M., and Prance, G. T. (1989). A comparison of the chemical composition of injured leaves in contrast to uninjured leaves of *Victoria amazonica* (Nymphaeaceae). *Ann. Bot.* 64, 697–706. doi: 10.1093/oxfordjournals.aob.a087896

Crovetto, R. N. M. (2012). Estudios etnobotánicos V. Nombres de plantas y su utilidad según los Mbya Guaraní de Misiones, Argentina. *Bonplandia* 21, 109–133. doi: 10.30972/bon.2121282

d'Orbigny (1835). *Voyage dans L'Amerique Meridoniale*. Paris: Pitois-Levrault, 289–290.

d'Orbigny (1840). Annales de Sciences Naturelles. *Bot. Ser.* 13:57.

Darling, A. C., Mau, B., Blattner, F. R., and Perna, N. T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* 14, 1394–1403. doi: 10.1101/gr.2289704

Dastpak, A., Osaloo, S. K., Maassoumi, A. A., and Safar, K. N. (2018). Molecular phylogeny of Astragalus sect. Ammodendron (Fabaceae) inferred from chloroplast ycf1 gene. *Ann. Bot. Fennici* 55, 75–82. doi: 10.5735/085.055.0108

de Lima, C. T., Machado, I. C., and Giulietti, A. M. (2021). Nymphaeaceae of Brasil. *Sitientibus série Ciências Biológicas* 21. doi: 10.13102/scb4986

de Queiroz, K. (1988). Systematics and the Darwinian revolution. *Philos. Sci.* 55, 238–259. doi: 10.1086/289430

de Queiroz, K. (1999). "The general lineage concept of species and the defining properties of the species category," in *Species, New Interdisciplinary Essays*, ed. R. A. Wilson (Cambridge, MA: MIT Press).

de Queiroz, K. (2007). Species concepts and species delimitation. *Systemat. Biol.* 56, 879–886. doi: 10.1080/10635150701701083

Decker, J. S. (1936). *Aspectos Biológicos da Flora Brasilieira*. São Leopoldo: BR Rotermund & Co, 640. doi: 10.5962/bhl.title.99988

Dong, W., Xu, C., Li, C., Sun, J., Zuo, Y., Shi, S., et al. (2015). *ycf1*, the most promising plastid DNA barcode of land plants. *Sci. Rep.* 5:8348. doi: 10.1038/srep08348

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.

Drescher, A., Ruf, S., Calsa, T. Jr., Carrer, H., and Bock, R. (2000). The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. *Plant J.* 22, 97–104. doi: 10.1046/j.1365-313x.2000.00722.x

Drummond, A. J., and Bouckaert, R. R. (2015). *Bayesian Evolutionary Analysis with BEAST*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139095112

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

Figueiredo, J., Hoorn, C., van der Ven, P., and Soares, E. (2009). Late Miocene onset of the Amazon River and the Amazon deep-sea fan: evidence from the Foz do Amazonas Basin. *Geology* 37, 619–622. doi: 10.1130/G25567A.1

Filipowicz, N., and Renner, S. S. (2012). *Brunfelsia* (Solanaceae): a genus evenly divided between South America and radiations on Cuba and other Antillean islands. *Mol. Phylogenet. Evolut.* 64, 1–11. doi: 10.1016/j.ympev.2012.02.026

Freudenstein, J. V., Broe, M. B., Folk, R. A., and Sinn, B. T. (2017). Biodiversity and the species concept—lineages are not enough. *Systemat. Biol.* 66, 644–656. doi: 10.1093/sysbio/syw098

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next- generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565

Gandolfo, M. A., Nixon, K. C., and Crepet, W. L. (2004). Cretaceous flowers of Nymphaeaceae and implications for complex insect entrapment pollination mechanisms in early angiosperms. *Proc. Natl. Acad. Sci. U.S.A.* 101, 8056–8060. doi: 10.1073/pnas.0402473101

Gaut, B. S., Muse, S. V., Clark, W. D., and Clegg, M. T. (1992). Relative rates of nucleotide substitution at the *rbcL* locus of monocotyledonous plants. *J. Mol. Evolut.* 35, 292–303. doi: 10.1007/BF00161167

Gickelhorn, R. (1966). *Thaddäus Haenkes Reisen und Arbeiten in Südamerika nach Dokumentarforschungen in Spanischen Archiven*. Wiesbaden: Franz Steiner Verlag.

GoogleEarth (2022). *The World's Most Detailed Globe*. Available Online at https://www.google.com/earth/index.html (accessed June 16, 2022)

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Gray, J. E. (1837). Description of *Victoria Regina* J.E. Gray. *Magaz. Zool. Bot.* 2, 440–442.

Gruenstaeudl, M., Nauheimer, L., and Borsch, T. (2017). Plastid genome structure and phylogenomics of Nymphaeales: conserved gene order and new insights into relationships. *Plant Systemat. Evolut.* 303, 1251–1270. doi: 10.1007/s00606-017-1436-5

Gutiérrez, P. A. G., Köhler, E., and Borsch, T. (2011). New species of *Buxus* (Buxaceae) from northeastern Cuba based on morphological and molecular characters, including some comments on molecular diagnosis. *Willdenowia* 43, 125–137. doi: 10.3372/wi.43.43115

Hanagarth, W. (1993). *Acerca de la Geoecologia de las Sabanas del Beni en el Noreste de Bolivia*. La Paz: Instituto de Ecologia.

He, D., Gichira, A. W., Li, Z., Nzei, J. M., Guo, Y., Wang, Q., et al. (2018). Intergeneric relationships within the early-diverging angiosperm family Nymphaeaceae based on chloroplast phylogenomics. *Int. J. Mol. Sci.*19:3780. doi: 10.3390/ijms19123780

Holway, T. (2013). *The Flower of Empire: The Ama'on's Largest Water Lily, the Quest to Make it Bloom, and the World it Helped Create*. Oxford: Oxford University Press.

Hooker, W. J. (1847). Victoria regia. *Curtis Bot. Magaz.* 73, 4275–4278.

Hoorn, C., Boschman, L. M., Kukla, T., Sciumbata, M., and Val, P. (2022). The Miocene wetland of western Amazonia and its role in Neotropical biogeography. *Bot. J. Lin. Soc.* 199, 25–35. doi: 10.1093/botlinnean/boab098

Hoorn, C., Wesselingh, F. P., Ter Steege, H., Bermudez, M. A., Mora, A., and Sevink, J. (2010). Amazonia through time: andean uplift, climate change, landscape evolution, andbiodiversity. *Science* 330, 927–931. doi: 10.1126/science.1194585

Hurrell, J. A., Puentes, J. P., and Arenas, P. M. (2016). Estudios etnobotánicos en la conurbación Buenos Aires-La Plata, Argentina: productos de plantas medicinales introducidos por inmigrantes paraguayos. *Bonplandia* 25, 43–52. doi: 10.30972/bon.2511270

Hutukara Associação Yanomami, and Associação Wanasseduume Ye'kwana (2020). *Scars in the Forest – The Growth of Illegal Mining in the Yanomami Indigenous Territory (YIL) in 2020. Report*. Available Online at: https://www.amazoniasocioambiental.org/en/radar/em-2020-garimpo-avancou-30-na-terra-indigena-yanomami-aponta-relatorio/ (accessed October 20, 2021)

Ibañez Montoya, M. V. (1984). *Trabajos Científicos y Correspondencia de Tadeo Haenke. La Expedicion Malaspina 1789-1794, Tomo IV*. Madrid: Lunwerg Editores.

Ikabanga, D. U., Stevart, T., Koffi, K. G., Monthe, F. K., Doubindou, E. C. N., 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. doi: 10.11646/phytotaxa.321.2.2

INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS (2021). *COORDENAÇÃO GERAL DE OBSERVAÇÃO DA TERRA. PROGRAMA DE MONITORAMENTO DA AMAZÔNIA E DEMAIS BIOMAS. Desmatamento -Ű Amazônia Legal.* Available online at: http://terrabrasilis.dpi.inpe.br/downloads/ (accessed October 20, 2021).

i-Terra (2021). *CIAT-Terra-I. Terra-i Perú*. Available online at: http://terra-i.org/terra-i/data/data-terra-i_peru.html (accessed October 20, 2021)

IUCN Standards and Petitions Committee (2019). *Guidelines for Using the IUCN Red List Categories and Criteria. Version 14. Prepared by the Standards and Petitions Committee*. Gland: IUCN Standards and Petitions Committee.

IUCN (2001). *IUCN Red List Categories and Criteria: Version 3.1. IUCN Species Survival Commission*. Gland: IUCN.

Jacobs, S. W., and Hellquist, C. B. (2011). New species, possible hybrids and intergrades in Australian *Nymphaea* (Nymphaeaceae) with a key to all species. *Telopea* 13, 233–243. doi: 10.7751/telopea20116016

Jin, J. J., Yu, W. B., Yang, J. B., Song, Y., DePamphilis, C. W., Yi, T. S., et al. (2020). GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol.* 21:241. doi: 10.1186/s13059-020-02154-5

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

Kinnup, V. F., and Lorenzi, H. (2014). "Unconventional food plants (PANC) in Brazil: identification guide, nutritional aspects and illustrated recipes," in *7th Brazilian Conference on Natural Product/ XXXIII RESEM Proceedings* (Sao Paulo), 2014.

Knoch, E. (1899). Untersuchungen über die morphologie, biologie und physiologie der blüte von *Victoria regia*. *Bibl. Bot.* 47, 1–60.

Korneliussen, T. S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* 15:356. doi: 10.1186/s12859-014-0356-4

Lamprecht, I., Schmolz, E., Blanco, L., and Romero, C. M. (2002). Energy metabolism of the thermogenic tropical water lily, *Victoria cruziana*. *Thermochim. Acta* 394, 191–204. doi: 10.1016/S0040-6031(02)00250-2

Langstroth Plotkin, R. (2012). Biogeography of the Llanos de Moxos: natural and anthropogenic determinants. *Geographica Helvetica* 66, 183–192. doi: 10.5194/gh-66-183-2011

Latorre, S. M., Lang, P. L., Burbano, H. A., and Gutaker, R. M. (2020). Isolation, library preparation, and bioinformatic analysis of historical and ancient plant DNA. *Curr. Protoc. Plant Biol.* 5:e20121. doi: 10.1002/cppb.20121

Lavin, L., and Pennington, R. T. (in press). "Non-monophyletic species are common in plants," in *Cryptic Species: Morphological Stasis, Circumscription, and Hidden Diversity*, eds A. K. Monro and S. J. Mayo (Cambridge: Cambridge University Press).

Les, D. H., Schneider, E. L., Padgett, D. J., Soltis, P. S., Soltis, D. E., and Zanis, M. (1999). Phylogeny, classification and floral evolution of water lilies (Nymphaeaceae; Nymphaeales): a synthesis of non- molecular, rbcL, matK, and 18S rDNA data. *Systemat. Bot.* 24, 28–46. doi: 10.2307/2419384

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

Lindley, J. (1837). *Victoria regia.* 1–5.

Lissambou, B. J., Couvreur, T. L., Atteke, C., Stévart, T., Piñeiro, R., Dauby, G., et al. (2019). Species delimitation in the genus *Greenwayodendron* based on morphological and genetic markers reveals new species. *Taxon* 68, 442–454. doi: 10.1002/tax.12064

Löhne, C., Borsch, T., and Wiersema, J. H. (2007). Phylogenetic analysis of Nymphaeales using fast- evolving and noncoding chloroplast markers. *Bot. J. Linn. Soc.* 154, 141–163. doi: 10.1111/j.1095-8339.2007.00659.x

Loureiro, J., Rodriguez, E., Doležel, J., and Santos, C. (2007). Two new nuclear isolation buffers for plant DNA flow cytometry: a test with 37 species. *Ann. Bot.* 100, 875–888. doi: 10.1093/aob/mcm152

Malme, G. O. A. (1907). Nagra anteckningar om *Victoria* Lindl., Sarskildt om *Victoria cruziana* D'Orb. *Acta Horti Berg.* 4, 3–16.

MapBiomas (2021). *MapBiomas Project- Collection Cobertura of the Annual Series of Land Use and Land Cover Maps of Brazil*. Available online at: https://plataforma.brasil.mapbiomas.org (accessed October 20, 2021).

Marengo, J. A., Cunha, A. P., Cuartas, L. A., Deusdara Leal, K. R., Broedel, E., Seluchi, M. E., et al. (2021). Extreme drought in the Brazilian Pantanal in 2019–2020: characterization, causes, and impacts. *Front. Water* 3:639204. doi: 10.3389/frwa.2021.639204

Mayo, S. J. (in press). "Cryptic species: a product of the paradigm difference between taxonomic and evolutionary species," in *Cryptic Species: Morphological Stasis, Circumscription, and Hidden Diversity*, eds A. K. Monro and S. J. Mayo (Cambridge: Cambridge University Press).

McAlvay, A. C., Armstrong, C. G., Baker, J., Elk, L. B., Bosco, S., Hanazaki, N., et al. (2021). Ethnobiology phase VI: decolonizing institutions, projects, and scholarship. *J. Ethnobiol.* 41, 170–191. doi: 10.2993/0278-0771-41.2.170

Meisner, J., and Albrechtsen, A. (2018). Inferring population structure and admixture proportions in low-depth NGS data. *Genetics* 210, 719–731. doi: 10.1534/genetics.118.301336

Mercado, J. (2021). *Tras el dorado. Crónicas de la explotación del oro en la Amazonía*. Cochabamba: La Libre.

Mereles, F., Céspedes, G., Soria, N., and de Arrúa, R. D. (2020). La importancia del trabajo botánico de Aimé Bonpland en Sudamérica y la incógnita de las colecciones botánicas realizadas en Paraguay. *Bonplandia* 29, 127–140. doi: 10.30972/bon.2924429

Monro, A. K. (in press). "Introduction," in *Cryptic Species: Morphological Stasis, Circumscription, and Hidden Diversity*, eds A. K. Monro and S. J. Mayo (Cambridge: Cambridge University Press).

Morales-Briones, D. F., Kadereit, G., Tefarikis, D. T., Moore, M. J., Smith, S. A., Brockington, S. F., et al. (2021). Disentangling sources of gene tree discordance in phylogenomic data sets: testing ancient hybridizations in Amaranthaceae sl. *Systemat. Biol.* 70, 219–235. doi: 10.1093/sysbio/syaa066

NASA (2021). *Earth Observatory. Deforestation*. Washington, DC: NASA.

Neubig, K. M., Whitten, W. M., Carlsward, B. S., Blanco, M. A., Endara, L., Williams, N. H., et al. (2009). Phylogenetic utility of ycf 1 in orchids: a plastid gene more variable than mat K. *Plant Systemat. Evolut.* 277, 75–84. doi: 10.1007/s00606-008-0105-0

Obermayer, R., Leitch, I. J., Hanson, L., and Bennett, M. D. (2002). Nuclear DNA C-values in 30 species double the familial representation in pteridophytes. *Ann. Bot.* 90, 209–217. doi: 10.1093/aob/mcf167

Ogilvie, H. A., Bouckaert, R. R., and Drummond, A. J. (2017). StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. *Mol. Biol. Evolut.* 34, 2101–2114. doi: 10.1093/molbev/msx126

Opitz, D. L. (2013). The sceptre of her pow'r': numphs, nobility, and nomenclature in early Victoria science. *Br. Soc. Hist. Sci.* 6–94. doi: 10.1017/S0007087413000319

Padial, J. M., and De la Riva, I. (2021). A paradigm shift in our view of species drives current trends in biological classification. *Biol. Rev.* 96, 731–751. doi: 10.1111/brv.12676

Pellegrini, M. O. O. (2020). *Nymphaeaceae, Flora do Brasil*. Rio de Janeiro: Jardim Botânico do Rio de Janeiro.

Pellicer, J., Garcia, S., Garnatje, T., Hidalgo, O., Korobkov, A. A., Dariimaa, S., et al. (2007). Chromosome counts in Asian Artemisia L.(Asteraceae) species: from diploids to the first report of the highest polyploid in the genus. *Bot. J. Linn. Soc.* 153, 301–310. doi: 10.1111/j.1095-8339.2007.00611.x

Pellicer, J., Kelly, L. J., Magdalena, C., and Leitch, I. J. (2013). Insights into the dynamics of genome size and chromosome evolution in the early diverging angiosperm lineage Nymphaeales (water lilies). *Genome* 56, 437–449. doi: 10.1139/gen-2013-0039

Pérez-Escobar, O. A., Bellot, S., Przelomska, N. A., Flowers, J. M., Nesbitt, M., Ryan, P., et al. (2021a). Molecular clocks and archeogenomics of a late period egyptian date palm leaf reveal introgression from wild relatives and add timestamps on the domestication. *Mol. Biol. Evolut.* 38, 4475–4492. doi: 10.1093/molbev/msab188

Pérez-Escobar, O. A., Dodsworth, S., Bogarín, D., Bellot, S., Balbuena, J. A., Schley, R. J., et al. (2021b). Hundreds of nuclear and plastid loci yield novel insights into orchid relationships. *Am. J. Bot.* 108, 1166–1180. doi: 10.1002/ajb2.1702

Pirie, M. D. (2015). Phylogenies from concatenated data: Is the end nigh? *Taxon* 64, 421–423. doi: 10.12705/643.1

Planchon, J. E. (1850). La Victoria regia au point de vue horticole et botanique: avec des observations sur la structure et les affinités des Nymphéacées. *Flore des Serres et des Jardins de l'Europe* 6: 193- 224, 249–254.

Planchon, J. E. (1851). La Victoria regia au point de vue horticole et botanique: avec des observations sur la structure et les affinités des Nymphéacées. *Flore des Serres et des Jardins de l'Europe* 7: 25–29, 49–53.

Poeppig, E. F. (1832). *Notizen aus dem Gebiete der Natur- und Heilkunde*. 35, 129–136.

Prance, G. T. (1974). *Victoria amazonica* ou *Victoria regia*? *Acta Amazonica* 4:6. doi: 10.1590/1809-43921974043005

Prance, G. T., and Arias, J. R. (1975). A study of the Floral Biology of *Victoria amazonica* (Poepp.). *Acta Amazonica* 5, 109–132. doi: 10.1590/1809-43921975052109

Renner, S. S. (2016). A return to Linnaeus's focus on diagnosis, not description: the use of DNA characters in the formal naming of species. *Systemat. Biol.* 65, 1085–1095. doi: 10.1093/sysbio/syw032

Robson, D. B., Wiersema, J. H., Hellquist, C. B., and Borsch, T. (2016). Distribution and ecology of a New Species of Water-lily, *Nymphaea loriana* (Nymphaeaceae), in Western Canada. *Can. Field Nat.* 130, 25–31. doi: 10.22621/cfn.v130i1.1787

Rodríguez-Rodríguez, P., de Paz, P. L. P., and Sosa, P. A. (2018). Species delimitation and conservation genetics of the Canarian endemic *Bethencourtia* (Asteraceae). *Genetica* 146, 199–210. doi: 10.1007/s10709-018-0013-3

Rosa-Osman, S. M., Rodrigues, R., de Mendonca, M. S., de Souza, L. A., and Piidade, M. T. F. (2011). Morfologia da flor, fruto e plantula de *Victoria amazonica* (Poepp.) J.C. Sowerby (Nymphaeaceae). *Acta Amazon.* 41, 21–28. doi: 10.1590/S0044-59672011000100003

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

Rutherford, S., Rossetto, M., Bragg, J. G., McPherson, H., Benson, D., Bonser, S. P., et al. (2018). Speciation in the presence of gene flow: population genomics of closely related and diverging *Eucalyptus* species. *Heredity* 121, 126–141. doi: 10.1038/s41437-018-0073-2

Scarpa, G. F. (2009). Wild food plants used by the indigenous peoples of the South American Gran Chaco: a general synopsis and intercultural comparison. *Angew. Bot.* 83, 90–101.

Scarpa, G. F., and Rosso, C. N. (2014). La etnobotánica moqoit inédita de Raúl Martínez Crovetto I: Descripción, actualización y análisis de la nomenclatura indígena. *Bol. Soc. Argent. Bot.* 49, 623–647. doi: 10.31055/1851.2372.v49.n4.9995

Schmidt, T. L., Jasper, M., Weeks, A. R., and Hoffmann, A. A. (2021). Unbiased population heterozygosity estimates from genome-wide sequence data. *Methods Ecol. Evolut.* 12, 1888–1898. doi: 10.1111/2041-210X.13659

Schneider, E. L. (1976). The floral anatomy of *Victoria* Schomb (Nymphaeaceae). *Bot. J. Linn. Soc.* 72, 115–148. doi: 10.1111/j.1095-8339.1976.tb01355.x

Schomburgk, R. H. (1837). Botanical society. *Athenaeum* 515:661.

Schubert, M., Lindgreen, S., and Orlando, L. (2016). AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. Notes* 9:88. doi: 10.1186/s13104-016-1900-2

Schultes, R. E. (1985). Several unpublished ethnobotanical notes of Richard Spruce. *Rhodora* 87, 439–441.

Schultes, R. E. (1990). Gifts of the amazon flora to the world. *Arnoldia* 50, 21–34.

Seton, M., Müller, R. D., Zahirovic, S., Gaina, C., Torsvik, T., and Shephard, G. (2012). Global continental and ocean basin reconstructions since 200 Ma. *Earth Sci. Rev.* 113, 212–270. doi: 10.1016/j.earscirev.2012.03.002

Seymour, R. S., and Mathews, P. G. D. (2006). The role of thermogenesis in the pollination biology of the amazon waterlily *Victoria amazonica*. *Ann. Bot.* 98, 1129–1135. doi: 10.1093/aob/mcl201

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351

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

Stevaux, J. C., Martins, D. P., and Meurer, M. (2009). Changes in a large regulated tropical river: the Paraná River downstream from the Porto Primavera Dam, Brazil. *Geomorphology* 113, 230–238. doi: 10.1016/j.geomorph.2009.03.015

Swenson, J. J., Carter, C. E., Domec, J. C., and Delgado, C. I. (2011). Gold mining in the Peruvian Amazon: global prices, deforestation, and mercury imports. *PLoS One* 6:e18875. doi: 10.1371/journal.pone.0018875

Templeton, A. R. (1989). "The meaning of species and speciation: a genetic perspective," in *The Units of Evolution: Essays on the Nature of Species*, eds D. Otte, and J. A. Endler (Sunderland, MA: Sinauer), 159–183.

The Global Biodiversity Information Facility [GBIF] (2022). *Free and Open Access to Biodiversity Data*. Available Online at: https://www.gbif.org (accessed June 16, 2022).

Thiers, B. (2016) *Index Herbariorum: A Global Directory of Public Herbaria and Associated Staff*. New York Botanical Garden's Virtual Herbarium. Available online at: http://sweetgum.nybg.org/science/ih/ (accessed May 5, 2022).

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E. S., Fischer, A., Bock, R., et al. (2017). GeSeq–versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* 45, W6–W11. doi: 10.1093/nar/gkx391

USGS Environmental Health Program (2019). *Mercury Isotope Ratios used to Determine Sources of Mercury to Fish in Northeast U.S. Streams*. Reston, VA: USGS.

Vásquez, G. C. (2015). "Indigenous people and climate change: causes of flooding in the Bolivian Amazon and consequences for the indigenous populaton," in *Inequality and Climate Change: Perspectves from the South*, ed. G. C. D. Ramos (Dakar: Council for the Development of Social Science Research in Africa (CODESRIA)), 121–136. doi: 10.2307/j.ctvh8r0w3.12

Warner, K. A., Rudall, P. J., and Frolich, M. W. (2008). Differentiation of perianth organs in nymphaeales. *Taxon* 57, 1096–1109. doi: 10.1002/tax.574006

Weissenborn, W. (1837). Notice respecting *Victoria regalis*. *Magaz. Natur. Hist. J. Zool. Bot. Mineral. Geol. Meteorol.* 1, 606–607.

Wells, T., Carruthers, T., Muñoz-Rodríguez, P., Sumadijaya, A., Wood, J. R., and Scotland, R. W. (2021). Species as a heuristic: reconciling theory and practice. *Systemat. Biol.* Online ahead of print, doi: 10.1093/sysbio/syab087

World Checklist of Vascular Plants (2022). *World Checklist of Vascular Plants, Version 2.0*. Kew: The Royal Botanic Gardens.

Xiang, Q. P., Wei, R., Zhu, Y. M., Harris, A. J., and Zhang, X. C. (2018). New infrageneric classification of Abies in light of molecular phylogeny and high diversity in western North America. *J. Systemat. Evolut.* 56, 562–572. doi: 10.1111/jse.12458

Yang, X. F., Wang, Y. T., Chen, S. T., Li, J. K., Shen, H. T., and Guo, F. Q. (2016). PBR1 selectively controls biogenesis of photosynthetic complexes by modulating translation of the large chloroplast gene *Ycf1* in *Arabidopsis*. *Cell Discov.* 2:16003. doi: 10.1038/celldisc.2016.3

Zhang, L., Chen, F., Zhang, X., Li, Z., Zhao, Y., Lohaus, R., et al. (2020). The water lily genome and the early evolution of flowering plants. *Nature* 577, 79–84. doi: 10.1038/s41586-019-1852-5

Zini, L. M., Galati, B. G., Gotelli, M., Zarlavcsky, G., and Ferrucci, M. S. (2019). Carpellary appendages in *Nymphaea* and *Victoria* (Nymphaeaceae): evidence of their role as osmophores based on morphology, anatomy and ultrastructure. *Bot. J. Linn. Soc.* 191, 421-439. doi: 10.1093/botlinnean/boz078

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

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

Copyright © 2022 Smith, Magdalena, Przelomska, Pérez-Escobar, Melgar-Gómez, Beck, Negrão, Mian, Leitch, Dodsworth, Maurin, Ribero-Guardia, Salazar, Gutierrez-Sibauty, Antonelli and Monro. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 14 July 2022
doi: 10.3389/fpls.2022.889988

A Bird's Eye View of the Systematics of Convolvulaceae: Novel Insights From Nuclear Genomic Data

Ana Rita G. Simões[1], Lauren A. Eserman[2]*, Alexandre R. Zuntini[1], Lars W. Chatrou[3], Timothy M. A. Utteridge[1], Olivier Maurin[1], Saba Rokni[1], Shyamali Roy[1], Félix Forest[1†], William J. Baker[1†] and Saša Stefanović[4†]*

[1] Royal Botanic Gardens, Kew, Richmond, United Kingdom, [2] Conservation & Research Department, Atlanta Botanical Garden, Atlanta, GA, United States, [3] Systematic and Evolutionary Botany Lab, University of Ghent, Ghent, Belgium, [4] Department of Biology, University of Toronto Mississauga, Mississauga, ON, Canada

OPEN ACCESS

Edited by:
Gerald Matthias Schneeweiss, University of Vienna, Austria

Reviewed by:
Gregory W. Stull, Kunming Institute of Botany (CAS), China
Richard Olmstead, University of Washington, United States
Charlotte Lindqvist, University at Buffalo, United States

***Correspondence:**
Ana Rita G. Simões
a.simoes@kew.org
Lauren A. Eserman
leserman@atlantabg.org

† These authors share senior authorship

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

Received: 04 March 2022
Accepted: 14 June 2022
Published: 14 July 2022

Citation:
Simões ARG, Eserman LA, Zuntini AR, Chatrou LW, Utteridge TMA, Maurin O, Rokni S, Roy S, Forest F, Baker WJ and Stefanović S (2022) A Bird's Eye View of the Systematics of Convolvulaceae: Novel Insights From Nuclear Genomic Data. Front. Plant Sci. 13:889988. doi: 10.3389/fpls.2022.889988

Convolvulaceae is a family of c. 2,000 species, distributed across 60 currently recognized genera. It includes species of high economic importance, such as the crop sweet potato (*Ipomoea batatas* L.), the ornamental morning glories (*Ipomoea* L.), bindweeds (*Convolvulus* L.), and dodders, the parasitic vines (*Cuscuta* L.). Earlier phylogenetic studies, based predominantly on chloroplast markers or a single nuclear region, have provided a framework for systematic studies of the family, but uncertainty remains at the level of the relationships among subfamilies, tribes, and genera, hindering evolutionary inferences and taxonomic advances. One of the enduring enigmas has been the relationship of *Cuscuta* to the rest of Convolvulaceae. Other examples of unresolved issues include the monophyly and relationships within Merremieae, the "bifid-style" clade (Dicranostyloideae), as well as the relative positions of *Erycibe* Roxb. and Cardiochlamyeae. In this study, we explore a large dataset of nuclear genes generated using Angiosperms353 kit, as a contribution to resolving some of these remaining phylogenetic uncertainties within Convolvulaceae. For the first time, a strongly supported backbone of the family is provided. *Cuscuta* is confirmed to belong within family Convolvulaceae. "Merremieae," in their former tribal circumscription, are recovered as non-monophyletic, with the unexpected placement of *Distimake* Raf. as sister to the clade that contains Ipomoeeae and *Decalobanthus* Ooststr., and Convolvuleae nested within the remaining "Merremieae." The monophyly of Dicranostyloideae, including *Jacquemontia* Choisy, is strongly supported, albeit novel relationships between genera are hypothesized, challenging the current tribal delimitation. The exact placements of *Erycibe* and *Cuscuta* remain uncertain, requiring further investigation. Our study explores the benefits and limitations of increasing sequence data in resolving higher-level relationships within Convolvulaceae, and highlights the need for expanded taxonomic sampling, to facilitate a much-needed revised classification of the family.

Keywords: *Ipomoea*, *Convolvulus*, *Cuscuta*, phylogeny, classification, Angiosperms353

INTRODUCTION

Convolvulaceae are a cosmopolitan plant family, widespread across tropical and temperate regions. It includes important food crops such as sweet potato (*Ipomoea batatas* L.) and water spinach (*Ipomoea aquatica* Forssk.) as well as a range of ornamental plants, including morning glories [e.g., *Ipomoea tricolor* Cav., *Ipomoea indica* (Burm.) Merr., *Ipomoea nil* (L.) Roth] and bindweeds (*Convolvulus* L. and *Calystegia* R. Br.). This family also harbors genus *Cuscuta* L. (dodders), members of which are stem parasites with little to no chlorophyll, representing one of 12 independent origins of haustorial parasitism across angiosperms (Nickrent, 2020). Convolvulaceae are phytochemically very diverse, containing a range of alkaloids, reflected in its wide array of phytotherapeutical and medicinal applications (Eich, 2008). The family is also well known to produce ergot alkaloids in the seeds of some species of *Argyreia* Lour. and *Ipomoea* L., as a result of an association with the fungus *Periglandula*, analogous to the infection of grasses by other related fungi (Beaulieu et al., 2021).

Convolvulaceae include 1,977 accepted species, classified into 60 genera and 12 tribes (Stefanović et al., 2003; Staples and Brummitt, 2007; POWO, 2022). The distribution of the species across genera and tribes is markedly uneven, with almost half of the diversity of the family being concentrated in tribe Ipomoeeae alone (835 species), followed by Convolvuleae (242 species), and Cuscuteae (218 species). These three tribes have been widely sampled in recent molecular phylogenetic studies (García et al., 2014; Williams et al., 2014; Muñoz-Rodríguez et al., 2019) and all three were shown to be monophyletic. However, the remaining nine tribes, containing 682 species distributed across 46 genera, are far less studied.

Classifications from before the onset of molecular phylogenetics and the application of monophyly as the primary principle of classification (Backlund and Bremer, 1998) lack the predictive power that phylogenetic classifications have, due to the absence of a connection between systematic arrangements and their shared, derived characters. In Convolvulaceae, as in many other plant families, the introduction of molecular phylogenetic approaches has brought novel perspectives on the classification systems, and has allowed the re-circumscription of higher-level taxonomic ranks (e.g., subfamilies, tribes, etc.). This improves the predictive power of the classification, and accelerates taxonomic progress.

The first family level phylogeny was produced based on plastid markers (*rbcL*, *atpB*, *psbE-J* operon, and *trnL-trnF* intron/spacer), sampling 52 of the 60 genera recognized at the time (Stefanović et al., 2002). Although the number of genera is currently the same, since this study there have been systematic changes, whereby some genera were synonymized with others, and new ones described (e.g., Staples, 2006; Buril et al., 2013, 2015; Simões and Staples, 2017; Petrongari et al., 2018; Simões and More, 2018; Simões et al., 2020; Staples et al., 2020) (**Supplementary Table 1**).

Stefanović et al. (2002) were the first to demonstrate the monophyly of Convolvulaceae, as well as of its three subfamilies as recognized at the time – Convolvuloideae, Humbertioideae, and Cuscutoideae – while exposing the non-monophyly of several tribes and genera across the family. Five of the tribes recognized at that time were clearly shown as non-monophyletic (Convolvuleae, Merremieae, Cresseae, Poraneae, Erycibeae), leading to a revision of the tribal classification that would reflect the newly recovered relationships (Stefanović et al., 2003). No subfamilies were formally proposed, but six large clades were identified which can be converted into informal subfamilies: Convolvuloideae, Dicranostyloideae (also called the "bifid style clade"), Cuscutoideae, Eryciboideae, Cardiochlamyoideae, and Humbertioideae. As for tribal ranking, 12 tribes were recognized, among which tribe "Merremieae" was the only one not confirmed to be monophyletic and left as of uncertain placement within the family (Stefanović et al., 2003).

A substantially expanded sampling of this group allowed for re-circumscription of its largest genus, *Merremia* Dennst. ex Endl. (Simões et al., 2015). It was found to be polyphyletic, and subsequently was divided into ten monophyletic genera, all of which are morphologically diagnosable and have received moderate-to-strong support in molecular phylogenies. However, while this study made significant progress in the generic circumscription within the tribe, the data and phylogenetic analyses were not robust enough to test the monophyly of the group itself, nor to demonstrate the relationships between the genera. Three of the genera were suggested to be sister to tribe Ipomoeeae (*Merremia* s.s., *Daustinia* Buril & A. R. Simões, and *Decalobanthus* Ooststr.), while the exact placement of the other seven have remained unresolved.

At the higher classification level, the "bifid style" clade, Dicranostyloideae, is the least resolved, with most of the relationships between its genera remaining uncertain. Also, their generic circumscription is questionable, with several of the genera (e.g., *Bonamia* Thouars, *Calycobolus* Willd. ex Schult.) having already been demonstrated not to be monophyletic (Stefanović et al., 2002). The same is true for *Jacquemontia*, a genus with a shortly divided style and ellipsoid stigmas, previously placed in tribe Convolvuleae (in the "single-style" clade, Convolvuloideae). As one of the most surprising results based on molecular phylogenetic evidence, this lineage has been moved to Dicranostyloideae, and assigned to its own tribe, Jacquemontieae (Stefanović et al., 2003). Interestingly, *Jacquemontia* seemed to be sister to the rest of Dicranostyloideae, but this relationship was only weakly supported (Stefanović and Olmstead, 2004) and has remained to be confirmed.

Finally, as one of the largest outstanding enigmas in Convolvulaceae, the phylogenetic position of *Cuscuta* within the family remains uncertain, owing largely to the rapid molecular evolution observed in this parasitic genus, across all three genomes, and accompanying analytical difficulties this entails (Stefanović and Olmstead, 2004). Based on the sequence data derived from all three plant genomes, at least two non-parasitic lineages are shown to diverge within the Convolvulaceae before *Cuscuta*. However, the exact sister group of *Cuscuta* could not be ascertained, even though many alternatives were rejected with confidence (Stefanović and Olmstead, 2004).

Recently, significant progress has been made in the incorporation of genomic data studies in tribes Ipomoeeae

(Eserman et al., 2014; Wu et al., 2018; Muñoz-Rodríguez et al., 2019) and Cuscuteae (Banerjee and Stefanović, 2019, 2020), with a clear benefit in providing a stronger framework for comparative work, e.g., estimating divergence times (Eserman et al., 2014; Carruthers et al., 2020), chromosome evolution (Ibiapino et al., 2022), etc. Higher-level organellomic approach, with an initial sampling across the family, uncovered some unusual scenarios of organellar evolution, especially regarding their mitogenomes, but the monophyly of "Merremieae" and the sister group of *Cuscuta* remained uncertain (Lin et al., 2022). However, most of the family has not yet caught up with these novel methodologies, which could finally bring clarity to the backbone relationships within Convolvulaceae, and unlock progress in systematics, biogeographic, and evolutionary studies at the level of the entire family.

A novel approach that has revolutionized evolutionary and systematic studies in the phylogenomic era is target sequence capture, in which genomic libraries are enriched for a specific set of genes (Dodsworth et al., 2019). In plants, the development of the Angiosperms353 (Johnson et al., 2019), a universal target capture probe set, has allowed standardization of genomic data generated for phylogenetic inference in angiosperms (Baker et al., 2021), enabling easier combination of different datasets and materials, including old herbarium specimens that are proven to be great source of genetic data (Brewer et al., 2019). This probe set has demonstrated its potential to advance phylogenetic studies and significantly resolve relationships with outstanding uncertainty across different taxonomic levels, from ordinal (e.g., Commelinales: Zuntini et al., 2021; Cornales: Thomas et al., 2021; Dipsacales: Lee et al., 2021) to familial (Orchidaceae: Eserman et al., 2021; Pérez-Escobar et al., 2021; Cyperaceae: Larridon et al., 2021) and even infra-generic levels (Shee et al., 2020; Slimp et al., 2021). Based on this universal probe set, large collaborative efforts such as Plant and Fungal Tree of Life Project (PAFTOL, Baker et al., 2022[1]) and Genomics for Australian Plants (GAP[2]) have generated an incomparable amount of genomic data for nearly all families and more than seven thousand genera of flowering plants, on which relationships among all plant families of angiosperms are being analyzed, at taxonomic and phylogenetic scales never attempted before. This unique and comprehensive dataset has an additional feature. The genes targeted by the Angiosperms353 probes are nuclear, which, in contrast to previous studies relying on plastid markers (Stefanović et al., 2002), is expected to facilitate the inclusion of parasitic species in broader analyses, given the documented gene loss tendency in plastid genomes, as observed in *Cuscuta* (Braukmann et al., 2013; Banerjee and Stefanović, 2019, 2020).

In the present study, we make use of these recently available nuclear genomic data towards: (1) exploring the deeper relationships between the main clades within Convolvulaceae; (2) testing the monophyly of tribes, insofar as the taxon sampling allows it; and (3) resolving the position of *Cuscuta* in relation to the remaining of Convolvulaceae.

[1]https://treeoflife.kew.org

[2]https://www.genomicsforaustralianplants.com/

MATERIALS AND METHODS

Taxon Sampling and Outgroup Selection

Genomic data of 34 out of the 60 genera of Convolvulaceae were analyzed, covering initially all 12 tribes (sensu Stefanović et al., 2003). A targeted effort was made to represent the main lineages in the family as informed by the molecular phylogenies of Stefanović et al. (2002) and Simões et al. (2015), as well as to include the most morphologically divergent genera. In our current study, half of the tribes are represented by a single species (Cuscuteae, Jacquemontieae, Aniseieae, Maripeae, Erycibeae, and Humbertieae). Therefore, the monophyly of these tribes cannot be assessed by our current data, but their relationships to each other and with other tribes within Convolvulaceae can be evaluated, commensurate with our "top-down" phylogenetic approach and focus of this paper. Other six tribes are represented by 2–8 species or genera allowing to assess not only their relationships with the remainder of the family but also their monophyly. Our sampling was limited to one species per genus except for the exceptionally large genus *Ipomoea*, which was represented with two species. Two other species within the family were excluded from further analysis: *Keraunea capixaba* Lombardi, for which preliminary analyses suggested a placement outside of Convolvulaceae; and *Anisea martinicensis* (Jacq.) Choisy, for insufficient data in comparison to the remaining samples. The only data available for *Convolvulus* is transcriptome data from the OneKP Project (One Thousand Plant Transcriptomes Initiative, 2019). Therefore, this genus was included in the analysis of exon data but excluded from supercontig analyses. Complete list of taxa and samples is provided in **Supplementary Table 2**.

Previous phylogenetic analyses have demonstrated that Convolvulaceae, including the parasitic genus *Cuscuta*, are monophyletic, and that *Humbertia madagascariensis* Lam. is sister to the rest of the family (Stefanović et al., 2002; Stefanović and Olmstead, 2004). We tested these hypotheses using a dataset of 35 species sampled from Convolvulaceae, five from the sister family Solanaceae, and one species from Montiniaceae, as the most distant outgroup in Solanales. Because alignments included individuals from three families, only exon sequences were used to estimate this phylogeny to improve homology assessment in alignment (**Supplementary Figure 1**).

DNA Extraction and Target Sequence Capture

Data production followed the protocol outlined in Baker et al. (2022). DNA samples were obtained from herbarium collections at Royal Botanic Gardens, Kew (K) or Kew's DNA Bank. DNA was isolated using the CTAB method (Doyle and Doyle, 1987), quantified using Quantus (Promega, Madison, WI, United States), and analyzed on agarose gels to assess the size distribution of DNA fragments. The ideal library size ranged between 350 and 450bp, and DNA samples with average fragments size above these thresholds were sonicated in a Covaris M220 Focused-ultrasonicator (Covaris, Woburn, MA, United States) prior to library preparation.

Dual indexed Illumina DNA libraries were prepared using NEBNext Ultra II DNA Library Prep Kit (New England Biolabs, Ipswich, MA, United States), following the manufacturer's protocol, with size selection not performed on highly degraded DNA samples. Libraries were amplified with 8–12 PCR cycles, depending on initial starting amount of DNA and later quantified using Quantus; library sizes were assessed using 4200 TapeStation and standard D1000 tapes (Agilent Technologies, Cheadle, United Kingdom). Up to 24 dual-indexed libraries were pooled in equimolar concentration prior to hybridization. Pooled libraries were hybridized with the Angiosperms353 probe set (Johnson et al., 2019; Arbor Biosciences myBaits Target Sequence Capture Kit) following manufacturer's protocol, ver. 4. Hybridization reactions were performed at 65°C for 24 h, followed by PCR amplification using NEBNext Q5 HotStart HiFi PCR Master Mix (New England BioLabs, Ipswich, MA, United States) and 12 cycles. Final hybridized pools were quantified and profiled as described above for individual libraries. Multiple enriched pools were combined, totaling up to 200 samples per sequencing lane, and sequenced using Illumina HiSeq at Macrogen (Seoul, South Korea). Raw Illumina reads have been deposited in the European Nucleotide Archive (PAFTOL BioProject PRJEB35285 and GAP Bioproject PRJEB49212).

Bioinformatic and Phylogenomic Analyses

Reads were first cleaned with Trimmomatic to remove barcode and adapter sequences and to remove reads with a quality score below 10 or reads less than 40 bp long (Bolger et al., 2014). Cleaned reads were assembled into genes using the HybPiper assembly pipeline (Johnson et al., 2016) using the expanded Angiosperms353 reference file (McLay et al., 2021). In short, this pipeline first maps reads to the reference file using bwa (Li and Durbin, 2009) and assembles reads into contigs using SPAdes (Bankevich et al., 2012; Prjibelski et al., 2020). Exonerate (Slater and Birney, 2005) is then used to align contigs to the target sequences in the reference file. Exons were then merged with introns to create "supercontigs," which are exon sequences with flanking introns recovered from the splash zone (Johnson et al., 2016). Supercontigs were created for each gene for each species. Because a whole genome was available for *Ipomoea triloba* (ASM357664v1; Wu et al., 2018), the Angiosperms353 sequences were recovered from the assemblies, using BLAST, as described in Baker et al. (2022). Sequences showing evidence of paralogy were removed from further analysis as has been done in previous phylogenetic analyses using target capture data (Wu et al., 2018; Eserman et al., 2021). There were very few paralog warnings in HybPiper assemblies, an average of 5 of 353 genes per sample, suggesting that paralogy is not a major issue in analyses of Angiosperms353 data in Convolvulaceae. Supercontig alignments for each one of the 353 nuclear loci were generated in PRANK (Löytynoja and Goldman, 2005, 2008; Löytynoja, 2014) and cleaned in Gblocks to remove positions with a gap in greater than 50% of individuals (Castresana, 2000; Talavera and Castresana, 2007). Alignment statistics were calculated with AMAS (Borowiec, 2016). Gblocks filtered alignments were then cleaned to remove samples without a sequence using a custom perl script – batch_Removeblank.pl.[3]

Phylogenetic trees were estimated using maximum likelihood methods in IQ-TREE 2 (Minh et al., 2020) based on a concatenated dataset. Clade support on gene trees was assessed using 1000 ultrafast bootstrap replicates as implemented in IQ-TREE 2 (Hoang et al., 2018). Coalescent analyses were performed in ASTRAL-III using default settings (Zhang et al., 2018), and clade supports were examined using multi-locus bootstrapping (Seo, 2008) and the clade polytomy tests (Sayyari and Mirarab, 2018).

One major goal of this study was to ascertain the placement of *Cuscuta* within the family. Owing to elevated substitution rates associated with photosynthesis loss, parasitic plants can not only be difficult to place in plant phylogeny in their own right, but their inclusion in analyses can also severely affect the resolution and support of other sampled, autotrophic taxa (e.g., Stefanović and Olmstead, 2004). To explore effects of inclusion of *Cuscuta* on our phylogenetic inference, analyses were conducted with two separate datasets: one with 34 species of Convolvulaceae including the functional outgroup *H. madagascariensis*, and the other containing the same taxa with addition of *Cuscuta australis* R. Br.

Alignments, gene trees, and unaligned gene datasets for the three taxon samples (Convolvulaceae and outgroups, Convolvulaceae without *Cuscuta*, and Convolvulaceae with *Cuscuta*) are available at: https://github.com/laeserman/Convolvulaceae_PAFTOL.

RESULTS

Between 309 and 351 genes of the total 353 genes were recovered for species within Convolvulaceae. Recovery was somewhat lower in outgroup species, with a range of 250-308 genes assembled per species. Within ingroup species, four genes had low recovery with a sequence assembled in only half or fewer of taxa (PAFToL Gene IDs 5354, 6886, 7013, and 7111). The mean sequence length per gene was 739 bp with gene assemblies ranging from 95 to 2796 bp (**Supplementary Table 3**). Exon alignments for the dataset containing Convolvulaceae and Solanaceae was 683 bp on average and ranged from 75 to 3127 bp in length; alignments contained on average 46% parsimony informative (PI) sites, with a range of 23% to 65%. Supercontig alignments for the dataset containing only Convolvulaceae species including *Cuscuta* were 1827 bp long on average and ranged from 206 to 7810 bp in length; alignments contained on average 25% PI sites, with a range of 13% to 34%. Supercontig alignments for the dataset containing Convolvulaceae without *Cuscuta* were 1774 bp long on average and ranged from 136 to 7567 bp; alignments contained on average 24% PI sites, with a range of 13% to 35% (**Supplementary Table 4**).

To assess the most suitable outgroup for further analysis of Convolvulaceae, we first included a full dataset of species in Convolvulaceae, sister family to Solanaceae, and more

[3]https://github.com/laeserman/Convolvulaceae_PAFTOL

distant outgroup from Montiniaceae. The species tree estimated in ASTRAL-III confirmed the monophyly of the family Convolvulaceae with 100% multilocus bootstrap support. As indicated in previous studies (Stefanović et al., 2002; Stefanović and Olmstead, 2004), *H. madagascariensis* was found here again to be sister to the rest of Convolvulaceae (**Supplementary Figure 1**). We recognize that in a larger analysis of Angiosperms353 data across all angiosperms, *Cuscuta* was recovered in this position instead of *Humbertia* (treeoflife.kew.org; Baker et al., 2022). It is likely that this broader analysis, using all plant families, was based on comparatively poor alignment due to the high rate of mutation in parasitic plant lineages, resulting in a spurious placement of *Cuscuta*. Because our focused analysis is showing *H. madagascariensis* as sister to the rest of Convolvulaceae, *H. madagascariensis* was used as the functional outgroup in further analyses of family level relationships, allowing for a greater inclusion of captured and assembled genes.

With this more inclusive gene dataset, we then assessed relationships in a sample focused only on Convolvulaceae species (**Figure 1**). Two datasets were generated, one including and another excluding *Cuscuta australis*. The trees resulting from coalescent analysis in ASTRAL-III with and without *Cuscuta* are similar in topology except for the placement of *Erycibe griffithii* C.B.Clarke. When *Cuscuta* is added to the analysis, it is found to have diverged after the diversification of Cardiochlamyeae and is sister to the rest of Convolvulaceae tribes including Ipomoeeae, Merremieae, Erycibeae, Cresseae, Dichondreae, Poraneae, Jacquemontieae, and Maripeae. The placement of *Cuscuta* within the family is supported by 82% multilocus bootstrap replicates in the ASTRAL-III analysis. This placement of *Cuscuta* is the same when the supercontig data are concatenated and analyzed in IQ-TREE 2, except that the support decreases to 59% (**Supplementary Figure 2**).

DISCUSSION

The analyses of nuclear genes targeted by the Angiosperms353 universal probes bring a fresh perspective on higher level relationships across Convolvulaceae. Some unanswered evolutionary and systematic questions remain, for which additional sampling will be required. We here provide an overview of the main novelties, hypotheses from previous studies that were confirmed, and outstanding challenges that will need to be addressed in the future.

Backbone of Convolvulaceae and Its Main Lineages

In most cases, the relationships between the larger clades and tribes are well supported (100% bootstrap) across the phylogenetic tree (**Figure 1** and **Supplementary Figure 2**). The most uncertain is the relationship between Cardiochlamyoideae and the clade that includes Cuscutoideae, Dicranostyloideae, and Convolvuloideae. While this does not impact the classification of the family at higher levels, it will likely hamper evolutionary studies that may aim to investigate processes beyond Dicranostyloideae and Convolvuloideae, and it is therefore an uncertainty that will need to be addressed in future studies, to enable any character evolution, biogeographic, and diversification analyses at the family level.

Humbertioideae and Eryciboideae are both lineages containing a single genus, even a monotypic genus in the case of *Humbertia*. This extraordinary species, a berry-producing tree endemic to Madagascar – *Humbertia madagascariensis* – was found to be sister to the rest of the family Convolvulaceae, as previous results had demonstrated (Stefanović et al., 2003). Placement of *Erycibe* as sister to the clade that contains Convolvuloideae and Dicranostyloideae (**Figure 1A**, In absence of *Cuscuta*) is also in agreement with earlier analyses; however, its position as sister to Convolvuloideae, in an analysis which includes *Cuscuta* (**Figure 1B**), is unexpected and without morphological support, to the best of the current knowledge of the family. We hypothesize that this is a methodological artifact, deriving from the addition of a representative of a potentially very long branched lineage (compare phylograms in **Supplementary Figure 2**).

Position of *Cuscuta*

The inclusion of *Cuscuta* in Convolvulaceae has, in early systematic studies, been under dispute, with authors having treated it as a separate family (Cuscutaceae) based on its parasitic life form (Dumortier, 1829; Roberty, 1952, 1964; Austin, 1973; Cronquist, 1988; Takhtajan, 1997). Molecular phylogenetic studies targeting the genus have demonstrated that it is, indeed, included in Convolvulaceae (Stefanović et al., 2002; Stefanović and Olmstead, 2004), although its exact position with respect to the remainder of the family has not been established with confidence. In the present study, we confirm the inclusion of *Cuscuta* within Convolvulaceae, with high support (99-100% bootstrap support; **Figure 1B** and **Supplementary Figure 2B**). However, the uncertainty of its position within the family remains. The ASTRAL-III analysis shows it as sister to the clade that includes Convolvuloideae, Dicranostyloideae, and *Erycibe*, with a moderate support (82% bootstrap; **Figure 1B**). This position is novel, and is not one of the potential placements seen in previous analyses (Stefanović and Olmstead, 2004). On the other side, the IQ-TREE 2 analysis of concatenated data places *Cuscuta* as sister to Convolvuloideae and Dicranostyloideae, with exclusion of *Erycibe*, but with weak support only (59% bootstrap; **Supplementary Figure 2B**). Concatenated data results (**Supplementary Figure 2**) also point to a likely reason behind the lack of support: a substantial substitution rate elevation and potential long-branch artifacts (Felsenstein, 1978; Bergsten, 2005) observed in many heterotrophic lineages (Nickrent, 2020), including *Cuscuta* (Stefanović and Olmstead, 2004). It is likely that with the addition of further taxon sampling across the family, and, in particular, the inclusion of additional species of *Cuscuta* to span the basal node of this divergent genus, relationships around *Cuscuta* can be resolved with greater support.

The monophyly of Cuscutoideae could also not be tested, because only one representative of this lineage was included in our current study. However, previous molecular phylogenetic analyses, with a broad sampling of ingroup taxa, have strongly

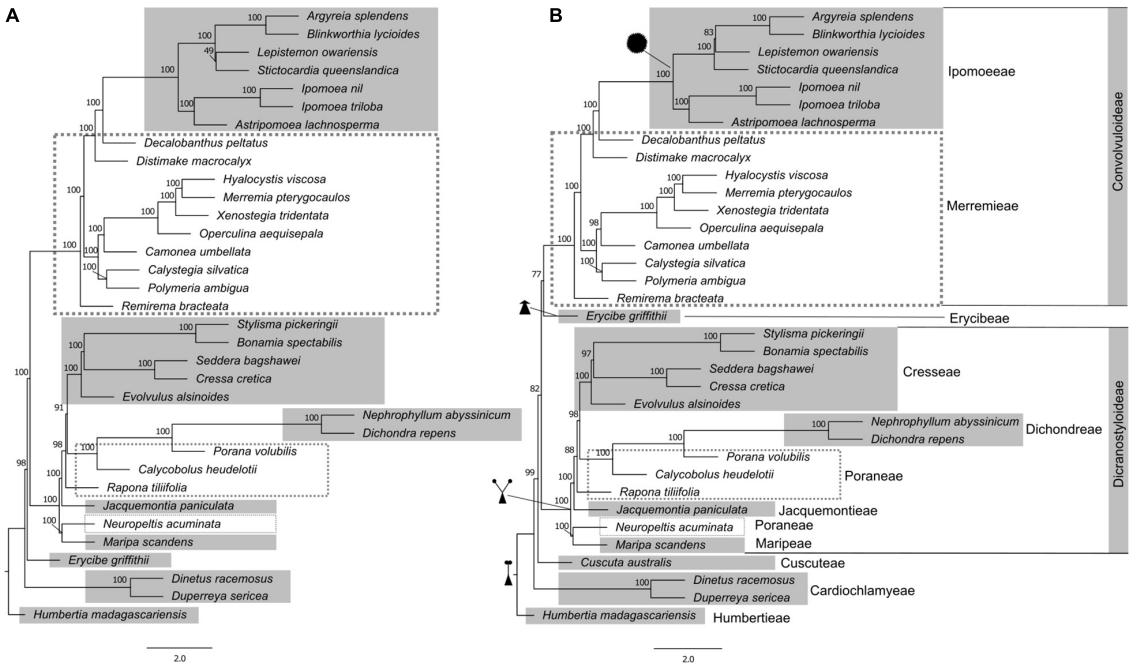

FIGURE 1 | Phylogeny of Convolvulaceae with different dataset composition estimated in ASTRAL-III, exploring the impact of the inclusion of *Cuscuta* in the coalescent analyses. **(A)** without *Cuscuta*; **(B)** with *Cuscuta*. Tribes are indicated, and known synapomorphies for the major clades are illustrated: single style, assumed plesiomorphic for the family; "bifid style", shared by the Dicranostyloideae clade including tribes Dichondreae, Poraneae, Jacquemontieae, and Maripeae; the lack of a style (sessile stigma) in Erycibeae; and spiny pollen, supporting tribe Ipomoeeae. Node support was assessed by multilocus bootstrapping in ASTRAL-III. ASTRAL-III only estimates internal branch lengths, which are in coalescent units; all terminal branch lengths are set to be equivalent.

demonstrated the monophyly of this subfamily (Stefanović et al., 2002; García et al., 2014). Such scenario is unlikely to change, considering the highly diverging morphology, ecology, and life form of the members of this subfamily, with respect to the rest of the family.

Monophyly and Circumscription of Dicranostyloideae

Most remarkably, our results offer strong support to the monophyly of one of the largest clades within the family, Dicranostyloideae, the "bifid style" clade. While the monophyly of Convolvuloideae received strong support across the board in previous molecular studies (Stefanović et al., 2002; Stefanović and Olmstead, 2004), this was not the case for Dicranostyloideae, where support varied from weak to moderate. One of the main issues was the unexpected position of *Jacquemontia*. This genus was once placed in tribe Convolvuleae based on morphology, but molecular phylogenetic analyses have suggested, albeit without substantial support, that it would belong in Dicranostyloideae instead (Stefanović et al., 2002; Stefanović and Olmstead, 2004). Our current results not only confirm the placement of *Jacquemontia* within Dicranostyloideae, sister to tribes Cresseae, Poraneae, and Dichondreae, but also suggests this relationship is very strongly supported (100% bootstrap support; **Figure 1**). Molecular phylogenetic studies by Stefanović et al. (2003) have proposed the circumscription of tribe Cresseae as including the following genera: *Bonamia*, *Cressa* L., *Evolvulus* L., *Hildebrandtia* Vatke (including *Cladostigma* Radlk. and *Sabaudiella* Chiov.), *Itzaea* Standl. & Steyerm., *Neuropeltis* Wall., *Neuropeltopsis* Ooststr., *Seddera* Hochst., *Stylisma* Raf., and *Wilsonia* G.L.Chu. Our phylogenetic analyses demonstrate with confidence that this tribe is not monophyletic in its current circumscription, due to the position of *Neuropeltis*, which is placed outside of any of the clades that include Poraneae, and sister to *Maripa* Aubl. In previous analyses, it was resolved as sister to the clade that included *Bonamia, Itzaea, Calycobolus, Dipteropeltis*, and *Rapona*, albeit with weak support (Stefanović et al., 2002). Poraneae, in its turn, are demonstrated to be paraphyletic, with *Rapona* Baill. being sister to the clade that includes Cresseae, Dichondreae, and part of Poraneae (*Porana* Burm. f. and *Calycobolus*).

Taken together, these are completely new insights into the relationships within Dicranostyloideae and these findings reinforce the importance of additional data in improving the analyses. Hence, these results are far from definitive, considering that at least ten genera of Dicranostyloideae still remain to be sampled, which could be key in clarifying at last the tribal delimitation within this clade.

Non-monophyly of "Merremieae" and Shifting Relationships to Ipomoeeae

Our analyses also confirm with great confidence the non-monophyly of the already dissolved "Merremieae", whose genera have been classified as "incertae sedis" (i.e., without tribal

placement) due to uncertainty regarding the delimitation of this tribe (Stefanović et al., 2003). The molecular phylogenetic analyses of Simões et al. (2015), which targeted this complex group in detail, suggested for the first time the placement of *Daustinia, Merremia,* and *Decalobanthus* within the clade that contains tribe Ipomoeeae with maximum support, as a grade, with each genus forming its own separate clade. The remainder of the tribe was resolved across multiple lineages, with a main larger clade that contained most of the genera of the "tribe" - *Distimake, Operculina* Silva Manso*, Camonea* Raf.*, Hewittia* Wight & Arn., *Hyalocystis* Hallier f.*, Xenostegia* D.F. Austin & Staples – and a number of unplaced species of *Merremia*. However, these results were generally doubtful, because most of the relationships were not significantly supported, which was attributed to the need of additional sequence data.

With our current data (**Figure 1** and **Supplementary Figure 2**), we have obtained a similar tree topology to that recovered previously for this group (Simões et al., 2015), but this time around the non-monophyly of "Merremieae" is for the first time demonstrated with high support. In addition, one completely novel relationship is presented: the largest genus previously segregated from *Merremia* s.l. – *Distimake* Raf. – is now actually found to be sister to the clade that includes *Decalobanthus* and Ipomoeeae (100% bootstrap).

Convolvuleae Nested Within "Merremieae"

An additional novel finding is the position of tribe Convolvuleeae within the "Merremieae", as sister to the largest "merremioid" clade, with strong support (100% bootstrap). Of all genera of Convolvuleae, *Convolvulus* was not sampled for our supercontig analysis, but was sampled for the exon analysis (**Supplementary Figure 1**), which supports its placement within the Convolvuleae and sister to *Calystegia*, consistent with prior results (Williams et al., 2014). This finding further supports the need to densely sample and reassess the tribal placement of the genera formerly included in "Merremieae", and their potential segregation into multiple monophyletic tribes.

CONCLUSION

In the history of classifications of Convolvulaceae, molecular phylogenetic analyses have been paramount to elucidate relationships within the family, where morphology was conflicting. As a result, new characters are arising as potentially taxonomically informative, and a new path for evolutionary and biogeographic hypotheses for the family is being illuminated.

The molecular phylogeny of Convolvulaceae by Stefanović et al. (2002) remains the most taxonomically comprehensive thus far. However, lack of support at the deeper relationships within the family hindered progress in systematics and evolutionary studies at higher taxonomic levels (subfamilies and tribes). Subsequent molecular phylogenetic studies have focused on smaller taxonomic groups within the family, and while informative, they have missed the taxonomic breadth that is necessary to fully resolve some of the outstanding problems within the family. A new age of genomic data, and the rise of large-scale collaborative projects that are globally fast-forwarding the sequencing of plant species, have provided an extraordinary amount of data, which we have accessed in this study as means to explore the deeper conflicts in the phylogeny of Convolvulaceae. The major clades within the family obtained in earlier phylogenetic studies seem strongly corroborated by the present analyses, although the tribal delimitation is still problematic due to the uncertainty of the classification of the "Merremieae" and the tribes within the Dicranostyloideae clade. The monophyly of several genera is also still to be further investigated (e.g., *Jacquemontia*, *Bonamia, Calycobolus*), which re-circumscription is likely to also have an impact in the tribal delimitation. While some key higher-level relationships are for the first time here consolidated, it is clear that a top-down reclassification of the family can now only be possible once this phylogenomic approach is expanded to additional taxa, with deeper sampling at generic and species level. Addition of key taxa in Convolvulaceae may lead to substantial taxonomic changes, paralleling those observed in Apiaceae (Clarkson et al., 2021) and Commelinaceae (Zuntini et al., 2021).

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. Raw Illumina reads can be found in the European Nucleotide Archive (PAFTOL BioProject: PRJEB35285 and GAP Bioproject: PRJEB49212). Alignments, gene trees, and unaligned gene datasets are available at: https://github.com/laeserman/Convolvulaceae_PAFTOL. The names of the repository/repositories and accession number(s) can be found in the article **Supplementary Material**.

AUTHOR CONTRIBUTIONS

AS coordinated the study and the writing of the manuscript, with collaboration of all co-authors. LE performed the analyses and contributed to manuscript writing. AZ, LC, and TU provided significant scientific contributions to the discussions and manuscript writing. OM and SRoy performed the laboratorial work. SRok collected samples. WB and FF conceived and supervised the generation of the genetic data as part of PAFToL project and contributed scientifically to the discussions and manuscript preparation. SS supervised the analyses, the scientific discussions, and the preparation of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was funded by a grant from the Calleva Foundation to the Plant and Fungal Trees of Life (PAFToL) Project at the Royal Botanic Gardens, Kew, and Natural Sciences and Engineering Research Council of Canada (Grant No. 326439) to SS. The Genomics for Australian Plants Framework Initiative was

supported by funding from Bioplatforms Australia (enabled by NCRIS) and partner organizations.

ACKNOWLEDGMENTS

We acknowledge the contribution of the Genomics for Australian Plants Framework Initiative consortium (https://www.genomicsforaustralianplants.com/consortium/) in the generation of data used in this publication. LE acknowledges the generous support of Atlanta Botanical Garden. AS was thankful to curators and taxonomists who have collected, identified and curated the herbarium specimens from which material was sampled in this study; also, staff at the Herbarium and Jodrell Laboratory of Royal Botanic Gardens, Kew who, through their continuous dedication, have provided in many ways the necessary conditions for this research to be performed. We also thank Dick Olmstead as well as three reviewers for their valuable inputs in earlier versions of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.889988/full#supplementary-material

Supplementary Figure 1 | Phylogeny of Convolvulaceae including outgroups from Solanaceae and Montiniaceae. **(A)** Tree estimated from a concatenated dataset of 349 genes analyzed in IQ-TREE2 with 1000 ultrafast bootstrap replicates. **(B)** Tree estimated in ASTRAL-III using gene trees from 349 gene trees. The main result from this analysis is that Convolvulaceae is monophyletic with 100% bootstrap support. Additionally, *Humbertia madagascariensis* is sister to the rest of the Convolvulaceae. Further analyses presented here use *H. madagascariensis* as the functional outgroup to improve alignment and thus tree estimation.

Supplementary Figure 2 | Phylogeny of Convolvulaceae with different dataset composition, exploring the impact of the inclusion of *Cuscuta* in the analyses. **(A)** without *Cuscuta*; **(B)** with *Cuscuta*. Both trees were estimated using a concatenated dataset of 349 genes that were aligned in PRANK and cleaned in Gblocks. The concatenated dataset was analyzed in IQ-TREE2 with 1000 ultrafast bootstrap replicates.

REFERENCES

Austin, D. F. (1973). The American Erycibeae (Convolvulaceae): *Capitalize Maripa*, *Dicranostyles*, and *Lysiostyles* I. Systematics. *Ann. Mo. Bot. Gard.* 60, 306–412. doi: 10.2307/2395089

Backlund, A., and Bremer, K. (1998). To Be or Not to Be. Principles of Classification and Monotypic Plant Families. *Taxon* 47, 391–400. doi: 10.2307/1223768

Baker, W. J., Bailey, P., Barber, V., Barker, A., Bellot, S., Bishop, D., et al. (2022). A Comprehensive Phylogenomic Platform for Exploring the Angiosperm Tree of Life. *Syst. Biol.* 71, 301–319. doi: 10.1093/sysbio/syab035

Baker, W. J., Dodsworth, S., Forest, F., Graham, S. W., Johnson, M. G., McDonnell, A., et al. (2021). Exploring Angiosperms353: an open, community toolkit for collaborative phylogenomic research on flowering plants. *Am. J. Bot.* 108, 1059–1065. doi: 10.1002/ajb2.1703

Banerjee, A., and Stefanović, S. (2019). Caught in action: fine-scale plastome evolution in the parasitic plants of *Cuscuta* section *Ceratophorae* (Convolvulaceae). *Plant Mol. Biol.* 100, 621–634. doi: 10.1007/s11103-019-00884-0

Banerjee, A., and Stefanović, S. (2020). Reconstructing plastome evolution across the phylogenetic backbone of the parasitic plant genus *Cuscuta* (Convolvulaceae). *Bot. J. Linn. Soc.* 194, 423–438. doi: 10.1093/botlinnean/boaa056

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. Comp. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021

Beaulieu, W. T., Panaccione, D. G., Quach, Q. N., Smoot, K. L., and Clay, K. (2021). Diversification of ergot alkaloids and heritable fungal symbionts in morning glories. *Commun. Biol.* 4:1362. doi: 10.1038/s42003-021-02870-z

Bergsten, J. (2005). A review of long-branch attraction. *Cladistics* 21, 163–193. doi: 10.1111/j.1096-0031.2005.00059.x

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

Braukmann, T. W., Kuzmina, M., and Stefanović, S. (2013). Plastid genome evolution across the genus *Cuscuta* (Convolvulaceae): two clades within subgenus *Grammica* exhibit extensive gene loss. *J. Exp. Bot.* 64, 977–989. doi: 10.1093/jxb/ers391

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

Buril, M. T., Simões, A. R., Carine, M., and Alves, M. (2013). *Austinia*, a new genus of Convolvulaceae from Brazil. *Phytotaxa* 186, 254–260. doi: 10.11646/phytotaxa.186.5.2

Buril, M. T., Simões, A. R., Carine, M., and Alves, M. (2015). *Daustinia*, a replacement name for *Austinia* (Convolvulaceae). *Phytotaxa* 197:60. doi: 10.11646/phytotaxa.197.1.8

Carruthers, T., Muñoz-Rodríguez, P., Wood, J. R. I., and Scotland, R. W. (2020). The temporal dynamics of evolutionary diversification in Ipomoea. *Mol. Phylogenet. Evol.* 146:106768. doi: 10.1016/j.ympev.2020.106768

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

Clarkson, J. J., Zuntini, A. R., Maurin, O., Downie, S. R., Plunkett, G. M., Nicolas, A. N., et al. (2021). A higher-level nuclear phylogenomic study of the carrot family (Apiaceae). *Am. J. Bot.* 108, 1252–1269. doi: 10.1002/ajb2.1701

Cronquist, A. (1988). *The Evolution And Classification Of Flowering Plants*. Bronx: The New York Botanical Garden.

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 in 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 from small quantities of fresh leaf tissue. *Phytochem. Bull.* 19, 11–15.

Dumortier, B.-C. (1829). *Analyse Des Plantes*. Paris: Tournay.

Eich, E. (2008). *Solanaceae and Convolvulaceae: Secondary Metabolites; Biosynthesis, Chemotaxonomy, Biological and Economic Significance; A Handbook*. Berlin: Springer. doi: 10.1007/978-3-540-74541-9

Eserman, L. A., Thomas, S. K., Coffey, E. E. D., and Leebens-Mack, J. H. (2021). Target sequence capture in orchids: developing a kit to sequence hundreds of single-copy loci. *Appl. Plant Sci.* 9:11416. doi: 10.1002/aps3.11416

Eserman, L. A., Tiley, G. P., Jarret, R. L., Leebens-Mack, J. H., and Miller, R. E. (2014). Phylogenetics and diversification of morning glories (tribe Ipomoeeae, Convolvulaceae) based on whole plastome sequences. *Am. J. Bot.* 101, 92–103. doi: 10.3732/ajb.1300207

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* 27, 401–410. doi: 10.2307/2412923

García, M. A., Costea, M., Kuzmina, M., and Stefanović, S. (2014). Phylogeny, character evolution, and biogeography of *Cuscuta* (dodders; Convolvulaceae) inferred from coding plastid and nuclear sequences. *Am. J. Bot.* 101, 670–690.

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

Ibiapino, A., García, M. A., Amorim, B., Baez, M. A., Costea, M., Stefanović, S., et al. (2022). The evolution of cytogenetic traits in *Cuscuta* (Convolvulaceae), the genus with the most diverse chromosomes in angiosperms. *Front. Plant. Sci.* 13:842260. doi: 10.3389/fpls.2022.842260

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:apps.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

Larridon, I., Zuntini, A. R., Léveillé-Bourret, E., Barrett, R. L., Starr, J. R., Muasya, A. M., et al. (2021). A new classification of Cyperaceae (Poales) supported by phylogenomic data. *J. Syst. Evol.* 59, 852–895. doi: 10.1111/jse.12757

Lee, A. K., Gilman, I. S., Srivastav, M., Lerner, A. D., Donoghue, M. J., and Clement, W. L. (2021). Reconstructing Dipsacales phylogeny using Angiosperms353: issues and insights. *Am. J. Bot.* 108, 1122–1142. doi: 10.1002/ajb2.1695

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

Lin, Y., Li, P., Zhang, Y., Akhter, D., Pan, R., Fu, Z., et al. (2022). Unprecedented organelle genomic variation in morning glories reveal independent evolutionary scenario of parasitic plants and the diversification of plant mitochondrial complexes. *BMC Biol.* 20:49. doi: 10.1186/s12915-022-01250-1

Löytynoja, A. (2014). "Phylogeny-aware alignment with PRANK," in *Multiple Sequence Alignment Methods*, ed. D. J. Russell (Totowa: Humana Press), 155–170. doi: 10.1007/978-1-62703-646-7_10

Löytynoja, A., and Goldman, N. (2005). An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl. Acad. Sci. U.S.A.* 102, 10557–10562. doi: 10.1073/pnas.0409137102

Löytynoja, A., and Goldman, N. (2008). Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* 320, 1632–1635. doi: 10.1126/science.1158395

McLay, T. G. B., Birch, J. L., Gunn, B. F., Ning, W., Tate, J. A., Nauheimer, L., et al. (2021). New targets acquired: improving locus recovery from the Angiosperms353 probe set. *Appl. Plant Sci.* 9:e11420. doi: 10.1002/aps3.11420

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., et al. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37, 1530–1534. doi: 10.1093/molbev/msaa015

Muñoz-Rodríguez, P., Carruthers, T., Wood, J. R. I., Williams, B. R. M., Weitemier, K., Kronmiller, B., et al. (2019). A taxonomic monograph of *Ipomoea* integrated across phylogenetic scales. *Nat. Plants* 5, 1136–1144. doi: 10.1038/s41477-019-0535-4

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

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

Pérez-Escobar, O. A., Dodsworth, S., Bogarín, D., Bellot, S., Balbuena, J. A., Schley, R. J., et al. (2021). Hundreds of nuclear and plastid loci yield novel insights into orchid relationships. *Am. J. Bot.* 108, 1166–1180. doi: 10.1002/ajb2.1702

Petrongari, F. P., Simões, A. R., and Simão-Bianchini, R. (2018). New combinations and lectotypifications in *Distimake* Raf. (Convolvulaceae). *Phytotaxa* 340, 297–300. doi: 10.11646/phytotaxa.340.3.12

POWO (2022). *Plants of the World Online*. London UK: Royal Botanic Gardens, Kew,

Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., and Korobeynikov, A. (2020). Using SPAdes de novo assembler. *Curr. Protoc. Bioinform.* 70:e102. doi: 10.1002/cpbi.102

Roberty, G. (1952). Genera Convolvulacearum. *Candollea* 14, 11–60.

Roberty, G. (1964). Les genres des Convolvulacées (esquisse). *Boissiera* 10, 129–156.

Sayyari, E., and Mirarab, S. (2018). Testing for polytomies in phylogenetic species trees using quartet frequencies. *Genes* 9:132. doi: 10.3390/genes9030132

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

Shee, Z. Q., Frodin, D. G., Cámara-Leret, R., and Pokorny, L. (2020). Reconstructing the Complex Evolutionary History of the Papuasian Schefflera Radiation Through Herbariomics. *Front. Plant Sci.* 11:258. doi: 10.3389/fpls.2020.00258

Simões, A. R., and More, S. (2018). Synopsis and lectotypification of *Distimake rhyncorhiza* (Dalzell) Simões & Staples (Convolvulaceae): a little known species from the Western Ghats (India). *Phytotaxa* 336, 293–298. doi: 10.11646/phytotaxa.336.3.8

Simões, A. R., and Staples, G. W. (2017). Dissolution of tribe Merremieae (Convolvulaceae) and a classification for its constituent genera. *Bot. J. Linn. Soc.* 183, 561–586. doi: 10.1093/botlinnean/box007

Simões, A. R., Culham, A., and Carine, M. (2015). Resolving the unresolved tribe: a molecular phylogenetic framework for Merremieae (Convolvulaceae). *Bot. J. Linn. Soc.* 179, 374–387. doi: 10.1111/boj.12339

Simões, A. R., Pisuttimarn, P., Pornpongrungrueng, P., and Chatrou, L. W. (2020). New combinations in *Decalobanthus* (Convolvulaceae). *Kew Bull.* 75:55. doi: 10.1007/s12225-020-09907-2

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

Slimp, M., Williams, L. D., Hale, H., and Johnson, M. G. (2021). On the potential of Angiosperms353 for population genomic studies. *Appl. Plant Sci.* 9:e11419. doi: 10.1002/aps3.11419

Staples, G. W. (2006). Revision of Asiatic Poraneae (Convolvulaceae) - *Cordisepalum*, *Dinetus*, *Duperreya*, *Porana*, *Poranopsis* and *Tridynamia*. *Blumea* 51, 403–491. doi: 10.3767/000651906X622067

Staples, G. W., and Brummitt, R. K. (2007). "Convolvulaceae," in *Flowering Plants of the World*, eds V. H. Heywood, R. K. Brummitt, A. Culham, and O. Seberg (Richmond Hill: Firefly Books), 108–110.

Staples, G. W., Simões, A. R., and Austin, D. F. (2020). A Monograph of Operculina (Convolvulaceae). *Ann. Mo. Bot. Gard.* 105, 64–138. doi: 10.3417/2020435

Stefanović, S., and Olmstead, R. G. (2004). Testing the Phylogenetic Position of a Parasitic Plant (Cuscuta, Convolvulaceae, Asteridae): bayesian Inference and the Parametric Bootstrap on Data Drawn from Three Genomes. *Syst. Biol.* 53, 384–399. doi: 10.1080/10635150490445896

Stefanović, S., Austin, D. F., and Olmstead, R. G. (2003). Classification of Convolvulaceae: a phylogenetic approach. *Syst. Bot.* 28, 791–806.

Stefanović, S., Krueger, L., and Olmstead, R. G. (2002). Monophyly of the Convolvulaceae and circumscription of their major lineages based on DNA sequences of multiple chloroplast loci. *Am. J. Bot.* 89, 1510–1522. doi: 10.3732/ajb.89.9.1510

Takhtajan, A. (1997). *Diversity And Classification Of Flowering Plants*. New York, NY: Columbia University Press.

Talavera, G., and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* 56, 564–577. doi: 10.1080/10635150701472164

Thomas, S. K., Liu, X., Du, Z.-Y., Dong, Y., Cummings, A., Pokorny, L., et al. (2021). Comprehending Cornales: phylogenetic reconstruction of the order using the Angiosperms353 probe set. *Am. J. Bot.* 108, 1112–1121. doi: 10.1002/ajb2.1696

Williams, B. R., Mitchell, T. C., Wood, J. R. I., Harris, D. J., Scotland, R. W., and Carine, M. A. (2014). Integrating DNA barcode data in a monographic study of *Convolvulus*. *Taxon* 63, 1287–1306. doi: 10.12705/636.9

Wu, S., Lau, K. H., Cao, Q., Hamilton, J. P., Sun, H., Zhou, C., et al. (2018). Genome sequences of two diploid wild relatives of cultivated sweetpotato reveal

targets for genetic improvement. *Nat. Commun.* 9:4580. doi: 10.1038/s41467-018-06983-8

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

Zuntini, A. R., Frankel, L. P., Pokorny, L., Forest, F., and Baker, W. J. (2021). A comprehensive phylogenomic study of the monocot order Commelinales, with a new classification of Commelinaceae. *Am. J. Bot.* 108, 1066–1086. doi: 10.1002/ajb2.1698

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

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

Copyright © 2022 Simões, Eserman, Zuntini, Chatrou, Utteridge, Maurin, Rokni, Roy, Forest, Baker and Stefanović. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

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: frontiersin.org/about/contact

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