Received: 28 September 2021 | Accepted: 26 May 2022

DOI: 10.1002/ajb2.16025

RESEARCH ARTICLE

American Journal of Botany

Incomplete lineage sorting and reticulate evolution mask species relationships in Brunelliaceae, an Andean family with rapid, recent diversification

José Murillo-A.[1] | Janice Valencia-D.[2] | Clara I. Orozco[1] | Carlos Parra-O.[1] | Kurt M. Neubig[2]

[1]Instituto de Ciencias Naturales, Universidad Nacional de Colombia, Carrera 30 # 45-03, edificio 425, Bogotá, D.C., Colombia

[2]School of Biological Sciences, Southern Illinois University Carbondale, 1125 Lincoln Dr., Carbondale, Illinois 62901-6509, USA

Correspondence

José Murillo-A., Instituto de Ciencias Naturales, Universidad Nacional de Colombia, Carrera 30 # 45-03, edificio 425, Bogotá D.C., Colombia.
Email: jcmurilloa@unal.edu.co

Abstract

Premise: To date, phylogenetic relationships within the monogeneric Brunelliaceae have been based on morphological evidence, which does not provide sufficient phylogenetic resolution. Here we use target-enriched nuclear data to improve our understanding of phylogenetic relationships in the family.

Methods: We used the Angiosperms353 toolkit for targeted recovery of exonic regions and supercontigs (exons + introns) from low copy nuclear genes from 53 of 70 species in *Brunellia*, and several outgroup taxa. We removed loci that indicated biased inference of relationships and applied concatenated and coalescent methods to infer *Brunellia* phylogeny. We identified conflicts among gene trees that may reflect hybridization or incomplete lineage sorting events and assessed their impact on phylogenetic inference. Finally, we performed ancestral-state reconstructions of morphological traits and assessed the homology of character states used to define sections and subsections in *Brunellia*.

Results: *Brunellia* comprises two major clades and several subclades. Most of these clades/subclades do not correspond to previous infrageneric taxa. There is high topological incongruence among the subclades across analyses.

Conclusions: Phylogenetic reconstructions point to rapid species diversification in Brunelliaceae, reflected in very short branches between successive species splits. The removal of putatively biased loci slightly improves phylogenetic support for individual clades. Reticulate evolution due to hybridization and/or incomplete lineage sorting likely both contribute to gene-tree discordance. Morphological characters used to define taxa in current classification schemes are homoplastic in the ancestral character-state reconstructions. While target enrichment data allows us to broaden our understanding of diversification in *Brunellia*, the relationships among subclades remain incompletely understood.

KEYWORDS

Angiosperms353, *Brunellia*, hybridization, incomplete lineage sorting (ILS), locus filtering, Oxalidales, phylogenetic discordance, rosids, target enrichment, trait evolution

Nuclear phylogenomic approaches using high-throughput sequencing methods (also known as next generation sequencing, e.g., Nowrousian, 2010; Morey et al., 2013; Heather and Chain, 2016; Kanzi et al., 2020), allow us to obtain data from the entire nuclear genome or a representative subset across the genome. Several approaches are available to obtain more low-copy nuclear data that have different costs and benefits (McKain et al., 2018). Trees inferred using large amounts of phylogenomic data permit inference of species trees with high support (e.g., from bootstrap analysis), but there can be underlying strong discordance between individual gene trees, and between gene trees and species trees (e.g., Maddison, 1997; Gadagkar et al., 2005; Kumar et al., 2012; Young and Gillung, 2020). It would therefore be useful to pinpoint and account for the processes underlying this incongruence.

Incongruence among phylogenies reconstructed from different data sets or genomes may be caused by various evolutionary processes that can act differently on individual gene trees, including hybridization or incomplete lineage sorting (ILS) due to rapid radiations (Maddison, 1997; Mao et al., 2019). The former reflects mixture of genetic material from different lineages (Rieseberg and Willis, 2007), whereas the latter implies retention of ancestral polymorphisms that can lead to disagreement among individual gene trees (Pamilo and Nei, 1988; Maddison, 1997). Both can mislead inference of species trees from gene trees. Several methods have been proposed to detect and account for these processes (e.g., Joly, 2012; Blischak et al., 2018; Lanfear, 2018), which improve our understanding of the dynamic processes that affect inference of relationships (e.g., Sun et al., 2015; Mao et al., 2019); reconciling gene-tree incongruences can also be used in the inference of a species tree (e.g., coalescence-based methods such as ASTRAL; Mirarab and Warnow, 2015).

Brunelliaceae is inferred to be a monophyletic family based on morphological evidence (Orozco, 2001). The family is a member of the order Oxalidales, in the rosid clade of eudicots. Its only genus, *Brunellia* Ruiz and Pav., has ~70 species, all of which are trees. Brunelliaceae are mostly represented in the Andes, with most species diversity found in Bolivia and Peru and north to Colombia and Venezuela; some extra-Andean species also grow in Mexico, Central America, the Greater Antilles, and the Guiana Shield. *Brunellia* is particularly rich in Colombia and some of its species reach high elevations (up to 3000 m) (Orozco et al., 2020).

Brunellia has been divided into two sections and 13 subsections (Cuatrecasas, 1985) or more recently into five sections and four subsections in a morphology-based phylogenetic study (Orozco, 2001). However, the monophyly of those taxonomic groups has not been tested to date using molecular characters. Our past attempts to find useful markers led us to assemble two plastomes in two species (*B. antioquensis* (Cuatrec.) Cuatrec. and *B. trianae* Cuatrec.) that were putatively distantly related, morphologically divergent, and classified in different sections of the genus (Valencia et al., 2020). However, as there was low variability (99.85% sequence similarity) between those plastomes, we decided to explore the use of nuclear loci here. Specifically, we analyzed nuclear genomic information obtained with the Angiosperms353 target sequence capture kit (Hyb-Seq; Weitemier et al., 2014), which captures hundreds of conserved single-copy protein-coding genes (Johnson et al., 2019), to address the following questions: (1) What phylogenetic relationships in *Brunellia* are inferred based on genomic data? (2) Are the morphological characters traditionally used to classify *Brunellia* useful for differentiating clades inferred using molecular data? And (3) To what degree does hybridization or incomplete lineage sorting (ILS) contribute to species diversification in *Brunellia*? While addressing these questions we also evaluated the utility of excluding poor quality loci based on criteria such as tree-to-tree distances, estimates of substitution saturation, and possible long-branch artifacts. We compared coalescent and concatenated species trees, providing insights into the degree to which ILS and hybridization may contribute to gene-tree discordance. Finally, we used this new phylogenetic framework to evaluate the evolution and homology of morphological traits thought to be important in infrageneric classification within *Brunellia*.

MATERIALS AND METHODS

Taxon sampling

We studied 53 species (out of ~70) in *Brunellia*, represented by 57 samples, and an additional eight outgroup taxa, including representatives of Celastrales, Malpighiales, and Oxalidales (Appendix S1). All data were newly generated in this study with the exception of *Cephalotus follicularis* Labill. [Cephalotaceae], which was retrieved from National Center for Biotechnology Information (website: http://www.ncbi.nlm.nih.gov/, accession PRJDB4484). We collected fresh leaf tissue from vouchered specimens in Bolivia, Colombia, Ecuador, and Peru, preserved in silica gel. Vouchers (Appendix S1) were deposited at the Herbario Nacional Colombiano (COL), at the Herbario Nacional de Bolivia La Paz (LPB) and at the Pontificia Universidad Católica del Ecuador Herbarium (QCA) (herbarium acronyms follow Thiers, 2020).

DNA extraction, library preparation and sequencing

We extracted DNA from dried tissues ground for 1 minute using zirconia beads in a Mini-BeadBeater 96 (Biospec Products, Bartlesville, Oklahoma, USA), followed by a standard CTAB protocol (Doyle and Doyle, 1987) modified by purifying the aqueous supernatant with silica columns (Neubig et al., 2014). We preserved total DNA in 1× Tris-EDTA (Fisher BioReagents BP2473-1; Thermo Fisher Scientific, Waltham, Massachusetts, USA) and standardized the DNA concentrations to 45-60 ng/μl, and then quantified them using a Qubit® 3.0 Fluorometer (Life Technologies, Carlsbad, California, USA). We examined DNA quality using agarose gel electrophoresis. Rapid Genomics LLC (Gainesville, Florida, USA) performed library preparation and sequencing, with DNA sheared using a sonicator to a mean fragment length of 400 bp. Fragments were end-repaired and A-tailed before the incorporation of unique dual-indexed Illumina adaptors. The libraries were then enriched using the Angiosperms353 target sequence capture kit (Arbor Biosciences, Ann Arbor, Michigan, USA) designed by Johnson et al. (2019), with sequencing performed on an Illumina HiSeqX (Illumina, Inc., San Diego, California, USA) to produce 150-bp, paired-end reads.

Data processing

We quality-trimmed raw reads using Trimmomatic version 0.39 (Bolger et al., 2014) to remove low-quality bases at the end and beginning of each read (when 4 bp windows had a quality score <Q20), and to remove reads shorter than 30 bp. After trimming, paired reads were processed using HybPiper version 1.3.1 (Johnson et al., 2016; available at website: https://github.com/mossmatters/Angiosperms353) with BWA mapper (Li et al., 2009) for aligning the reads to the DNA targets, and SPAdes (Bankevich et al., 2012) for de novo assembly of reads. We performed the first capture using the "Angiosperms353_targetSequences" fasta file available on the HybPiper website. We recovered exonic regions with the "exonerate" script. The sample "*Brunellia inermis* Ruiz and Pav., *Orozco 4085*" was selected due to its high coverage to create a new customized set of exon data targets, to maximize data recovery in the family. We recovered exon data and supercontigs (exons + introns) of the 353 genes using the "reads_first.py" and "exonerate_hits.py" scripts. To visualize recovery efficiency, we summarized exon coverage using R version 3.6.3 (R Core Team, 2020) with the "gen_recovery_heatmap_ggplot.R" script (available at website: https://github.com/mossmatters/HybPiper). We retrieved the DNA sequences of *C. follicularis* (GenBank PRJDB4484) from NCBI for the 353 protein targets with the format option: 'fasta CDS,' and included them in the individual gene datasets with a custom Python script (Williams, 2022). We inspected putative paralogs with the "paralog_investigator.py" script. Each gene with paralogs was retrieved using the "paralog_retriever" script and GNU parallel (Tange, 2018), and the copies were aligned using MAFFT version 7.450 (Katoh and Standley, 2013). The phylogenetic relationships depicted by each gene were reconstructed with FastTree version 2.1.11 (Price et al., 2010) plugins in Geneious Prime version 2020.0.3 (website https://www.geneious.com) to understand the nature of putative paralogous cases.

Loci filtering

We evaluated ten parameters for each gene in the exon and supercontigs datasets, to minimize potential bias produced by a strong, but misleading signal (Shen et al., 2017), such as sequence saturation, long branch attraction (Felsenstein, 1978; Hendy and Penny, 1989), and potential hidden paralogy due to polyploidization (Wolfe, 2001; Veitia, 2005) (Table 1). Genes that were over or under the 1.5 interquartile range (see Table 1 for each case) for any of the considered parameters were excluded from all downstream analyses. We called the exons dataset without paralogs and loci with potential biases "EWE" (exons with exclusions) and the supercontigs dataset with the same exclusions "SCWE" (supercontigs with exclusions). Additionally, the phylogenetic support of the datasets with exclusions (for both EWE and SCWE) was calculated with the average bootstrap support using TreSpEx version 1.1 (Struck, 2014). This measure was obtained using all branches of the best maximum likelihood (ML) tree of each gene tree. The averages were depicted in a density plot to visualize the number of genes with an average >60% of bootstrap support in each dataset using R version 4.0.2 (R Core Team, 2020).

Phylogenetic analyses

We analyzed the four datasets of exons and supercontigs, with and without genes excluded (based on the filtering step above), on the Galaxy platform using Osiris phylogenetic tools (website: https://galaxyproject.org/; Oakley et al., 2014).

TABLE 1 Parameters used for reducing phylogenetic bias.

Parameter	Cause of bias in the phylogenetic analysis	Description	Tail of the distribution that defines the region of gene outliers
1. Relative likelihood	Phylogenetic contribution bias	Relative contribution of each gene to the total likelihood score	$Q_3 + 1.5(IQR)$
2. R^2 of the linear regression	Saturation	Linear regression between patristic and uncorrected pairwise distances for each gene	$Q_1 - 1.5(IQR)$
3. Slope of the linear regression			$Q_1 - 1.5(IQR)$
4. Long branch score upper quartile	Long branch attraction	Which are based on averages and speed of evolutionary rates. These parameters were calculated using TreSpEx version 1.1	$Q_1 - 1.5(IQR)$
5. Long branch score heterogeneity			$Q_3 + 1.5(IQR)$
6. Tip to root upper quartile			$Q_3 + 1.5(IQR)$
7. Tip to root heterogeneity			$Q_3 + 1.5(IQR)$
8. Average patristic differences		As a proxy for genes affected by long-branch attraction	$Q_3 + 1.5(IQR)$
9. Matching Splits distance	Potential paralogy	It corresponds to the distance of each gene tree to the ML tree of supercontigs. These parameters were calculated through the platform https://eti.pg.edu.pl/treecmp (Bogdanowicz et al., 2012)	$Q_3 + 1.5(IQR)$
10. Robinson-Foulds distance			$Q_3 + 1.5(IQR)$

We used MAFFT (Katoh and Standley, 2013) to produce alignments. To avoid discrepancies between sequences and samples with a high number of gaps, the alignments were trimmed with TrimAl version 1.4 using the command "automated1", which selects the best method for trimming according to data characteristics (Capella-Gutierrez et al., 2009). We performed phylogenetic analyses using loci concatenation and multispecies-coalescent methods. For the first one, we analyzed concatenated alignments using maximum likelihood (ML) in IQ-TREE version 2.0.6 (Minh, Schmidt et al., 2020). The best substitution model was selected using ModelFinder under BIC (Kalyaanamoorthy et al., 2017). We used Ultrafast bootstraps (UFBoot) with 1000 replicates to calculate branch support (Hoang et al., 2018). To quantify the genealogical concordance within each dataset, we used the gene concordance factor (gCF, the proportion of inferred gene trees that contain that branch) and the site concordance factor (sCF, the proportion of inferred sites supporting a branch in a given tree). Additionally, we inspected the ML trees produced from the EWE and SCWE datasets to determine the proportion of gene trees (gDF1, gDF2) and sites (sDF1, sDF2) that support the two main alternative topologies (Minh, Hahn et al., 2020). We also used IQ-TREE to build the trees based on individual loci that were subsequently included in ASTRAL analyses (Zhang et al., 2018). For each branch in the species tree, we recovered the local posterior probability (LPP) support, which measures the probability of the branch given the dataset, and the Quartet support percentage (QS), which measures the conflict of gene trees for each species-tree branch (Zhang et al., 2018).

We also performed gene tree-discordance detection, network analysis, hybridization test, ILS evaluation, and the ancestral state reconstruction of morphological characters using the SCWE dataset. When a phylogenetic topology was necessary for a test, we used the phylogeny obtained with the multi-coalescent method on that dataset. This particular topology was selected for three reasons: (1) coalescent analyses of multiple nuclear genes with independent segregation capture the evolutionary history of the group better than concatenated analysis (e.g., Mirarab and Warnow, 2015); (2) the supercontig dataset is more phylogenetically informative than the exon dataset in this study; and (3) the exclusion of loci reduced the effect of biased data.

Gene-tree discordance detection

We evaluated the predicted extent of gene duplication, estimating the "internode certainty all" (ICA) support (Salichos et al., 2014) and the number of genes that support each branch in the SCWE Astral tree, using Phyparts (Smith et al., 2015). The individual SCWE dataset gene trees used in the analysis were rooted on *Monteverdia ebenifolia* (Reissek) Biral [Celastraceae] using the STRAW webserver (http://bioinformatics.publichealth.uga.edu/SpeciesTreeAnalysis/index.php). Gene trees were also visualized with DensiTree version 2.0.1 (Bouckaert, 2010). To identify possible clusters of trees that might indicate the prevalence of two or more topologies, we also made a multidimensional scaling visualization (MDS) of the gene trees. In the MDS, previously calculated Robinson-Foulds tree-to-tree distances (Robinson and Foulds, 1981) on the SCWE dataset were depicted using R version 4.0.2 (R Core Team, 2020).

Network analysis

A possible phylogenetic conflict within the SCWE dataset was assessed using SplitsTree4 version 4.16.2 (Huson and Bryant, 2006) which depicts the incompatible phylogenetic signal as a web that connects the species. The input matrix was made by the concatenation of alignments of the mentioned dataset using FasconCat-G version 1-04 (Kück and Longo, 2014). An unreduced median network was generated using MedianNetwork, a method that uses all sites that have exactly two different states and that excludes missing states and gaps.

Hybridization and ILS detection

We detected probable introgression within *Brunellia* using the HyDe (Blischak et al., 2018) program, which recognizes hybridization events using phylogenetic invariants within a model that includes coalescence and hybridization. HyDe evaluates all possible groups of four samples and assigns one as the outgroup, two as parental populations (P1 and P2), and a hybrid produced from a mixture of P1 and P2. The program evaluates the presence of introgression and the probable contribution of P2 and P1 (called γ and 1-γ, respectively). We conducted an exhaustive test among phylogenetic terminals using a concatenated alignment of the SCWE dataset using the python script "run_hyde.py" with the "–ignore_amb_sites" flag. Tests with significant *P*-values after Bonferroni corrections (γ between 0 and 1, and Z-score >3) were considered strong evidence of hybridization.

We evaluated ILS using a method developed by Lanfear (2018) in which the null hypothesis (H_0) of having approximately the same number of genes or sites supporting the two alternative topologies is tested. Following Huson et al. (2005), the H_0 scenario is the result of independent occurrences of lineage sorting, whereas the alternative (H_1) scenario involves reticulation. We employed a chi-square test using an R script due to Lanfear (2018), to identify branches with significant differences among genes (gDF1, gDF2) or sites (sDF1, sDF2).

Character evolution

We selected seven morphological characters used previously in traditional taxonomic treatments to define sections and subsections by Cuatrecasas (1970, 1985) and Orozco (2001).

To delimit character states, we considered herbarium specimens and data from the Orozco study (2001) (Appendix S2), with terminology based on Orozco (2001). We evaluated leaf complexity, inflorescence complexity, fertile portion of the inflorescence, calyx merosity, ratio of carpel to calyx merosity, carpel number, and endocarp shape. Ancestral character states were reconstructed using Mesquite version 3.51 (Maddison and Maddison, 2018) under ML methods and the Markov k-state one-parameter (Mk1 model), which assumes equal rates of change among character states (Lewis, 2001). The asymmetric model (Mk2 model) was also evaluated, which allows a different rate between gains and losses (Pagel, 1999). The best model was selected using the Asymmetry Likelihood Ratio Test option included in Mesquite version 3.51 (Maddison and Maddison, 2018).

RESULTS

Data assembly

We generated Hyb-Seq data for 57 samples of Brunelliaceae and for eight outgroups representing closely related families from the Oxalidales, Celastrales, and Malpighiales (Appendices S3 and S4). On average, we recovered 4,425,973 reads per sample (SD = 1,653,701), of which 31% mapped to the targeted loci. Out of the total 353 targeted loci, we obtained data from 350 genes with an average of 62.9 samples per gene (SD = 4.2). Twenty-seven genes with gene duplications were detected in our analyses (Appendix S5). Of those, nine were identified as paralogous duplications within *Brunellia* and were removed from the downstream analyses. The other 18 genes correspond to gene duplications within the outgroup samples, so we kept one copy selected through the pipeline based on high coverage, depth and similarity with the reference. We removed the paralogues identified using HybPiper scripts (Johnson et al., 2016; available at website: https://github.com/mossmatters/Angiosperms353) to obtain 337 exon loci and 337 supercontig loci.

Locus exclusion

We excluded 35 genes from the exon dataset and 59 genes from the supercontigs dataset based on the evaluation of ten parameters (Table 1), and created two reduced datasets (EWE and SCWE datasets). The excluded genes correspond to the outliers of the distribution of one or multiple parameters (see full list in Appendices S6 and S7) in the exon and supercontigs datasets. This exclusion increased the support values of the branches in some but not all comparative analyses. When all datasets (exons and supercontigs with and without exclusions) were analyzed with coalescence and concatenation approaches, eight different topologies were recovered (Figures 1 and 2; Appendix S8).

We used the average ultrafast bootstrap (UFBoot) support of each gene tree to evaluate the phylogenetic support for the EWE and SCWE datasets. Gene trees inferred from the supercontigs dataset had higher average values of UFBoot support than the trees inferred from the exons dataset (Appendix S9). In the supercontigs dataset, only 22 gene trees had values less than 60%, while in the exons dataset 234 gene trees had averages under 60%. After removing paralogs and pruning with TrimAl (but before excluding loci with potential biases), the exon dataset included 302 genes (with 212,329 sites), and the supercontigs dataset included 317 genes (with 850,115 sites) (Table 2). The supercontigs dataset had a higher number of informative sites and indels than the exon dataset (Table 2).

Phylogenetic analyses

All the analyses recovered two clades with strong support (BS = 100, LPP = 1, Figures 1 and 2; Appendix S8) within Oxalidales. One consisted of Connaraceae and Oxalidaceae, and the other of Brunelliaceae, Cephalotaceae, Cunoniaceae, and Elaeocarpaceae. The relationships among these last four families were variable and depended on the dataset used. However, Cephalotaceae was always sister to the other three families, and in most of our analyses, Brunelliaceae was sister to Cunoniaceae + Elaeocarpaceae (Figure 1). Some of the topologies obtained with the exon dataset showed Cunoniaceae sister to Brunelliaceae + Elaeocarpaceae (Appendix S8).

Brunellia (and therefore Brunelliaceae) is inferred to be monophyletic and comprises two major well-supported clades (BS = 100, LPP = 0.99-1) (Figures 1 and 2; Appendix S8). The phylogenetic analyses using concatenation and coalescence methods on the complete datasets and on the datasets with exclusions (EWE and SCWE) depicted similar general patterns, but none were entirely identical (Figure 1, Appendix S8). We always recovered clades I and II, with low gCF and sCf support (Appendices S10 and S11), moderate QS values, and short branches compared with the remaining Oxalidales (Appendices S10, and S11, S12). Within clade I, there were six lineages. Of these, clade Ia, comprising *B. brunnea* J.F. MacBr., *B. coroicoana* Cuatrec., and *B. rhoides* Rusby, was sister to all other species. Each of these six lineages had high support; however, relationships among them were not consistent (Figures 1 and 2; Appendix S8). Relationships within each lineage were consistent, except in subclade Ia. Subclade Ic was not recovered when the EWE dataset was analyzed by coalescence (Figure 2E).

Seven main lineages were recovered within clade II (Figures 1 and 2; Appendix S8). *Brunellia acostae* Cuatrec. was sister to all other species in this clade. Relationships among and within the remaining subclades were variable. However, the relationships among the five species in subclade IIb were consistent in all analyses and the relevant branches were most often highly supported. The position of *Brunellia hygrothermica* Cuatrec. was variable within clade II, appearing sister to different subclades in the various

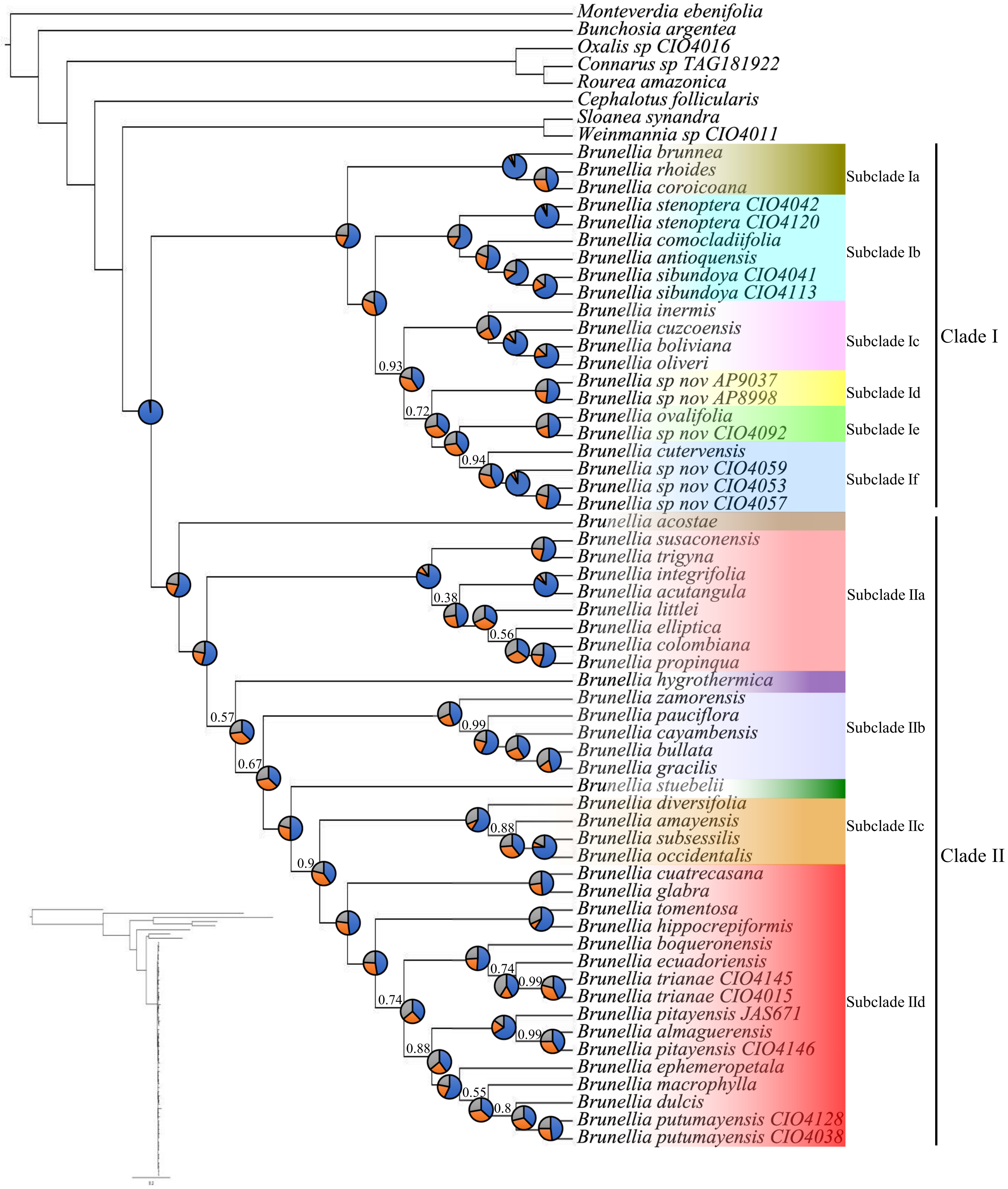

FIGURE 1 ASTRAL-based species tree of Brunelliaceae based on Angiosperms353 of the SCWE dataset (supercontigs excluding paralogs and loci with potential bias). Numbers above branches indicate local posterior probability (LPP) values. Pie charts show quartet support for the main topology (blue), the first alternative (orange), and the second alternative (gray). Insert corresponds to ML phylogram of the same dataset.

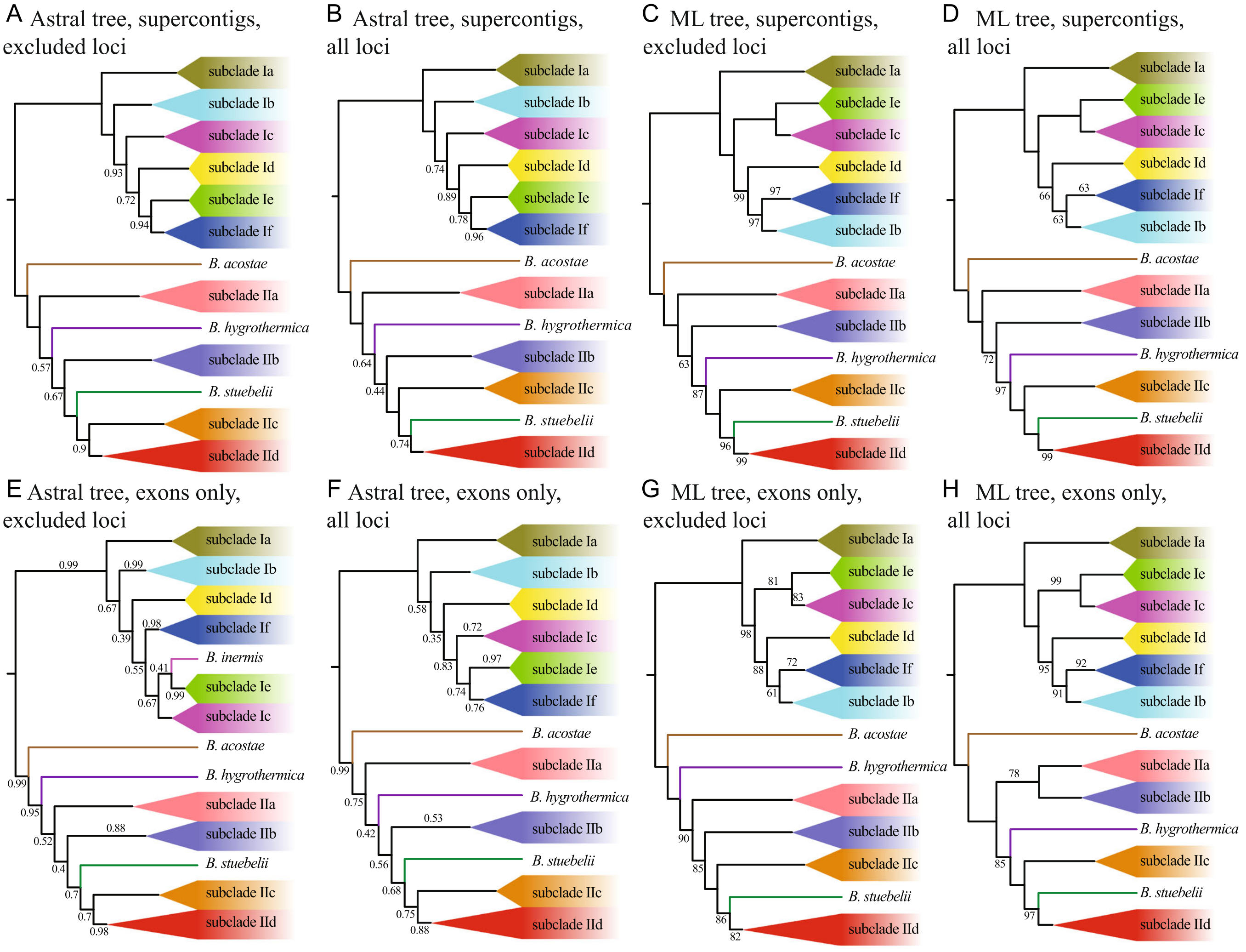

FIGURE 2 Species tree of Brunelliaceae based on exons and supercontigs using data-set with and without locus exclusion. Trees were inferred using ASTRAL and maximum likelihood analysis of concatenated data. Each subclade has been condensed to show the variable relationships within the subclades. Numbers below and above branches are LPP values for trees obtained by coalescence and ultrafast bootstrap for trees obtained with maximum likelihood.

analyses. Finally, the clade formed by *B. stuebelii* Hieron. + subclade IIc + subclade IId was recovered in all analyses.

Gene-tree discordance

Strong incongruence was detected among gene trees in the four analyzed datasets (exon, supercontigs, with and without exclusions), exemplified by eight different topologies (Figures 1 and 2; Appendix S8). Despite recovering generally high branch support, a high degree of locus discordance in the four datasets was found according to the low values of gCF, sCF, and QS for the main topology (Figure 1, Appendix S8). The backbone of clades I and II had the lowest gCF values related to the most terminal nodes, and high values of gDFP that indicate high discordance due to polyphyly (Figure 3, Appendix S13).

When the SCWE dataset was evaluated on the coalescent tree, a high discordance was found for branches of the backbone of clades I and II (Figure 1). The most concordant nodes among the gene trees were the most recent common ancestor of *Brunellia* and some terminal nodes that support groups of few species (i.e., subclades). Additionally, the ICA values were close to zero or below (Figure 4), indicating that there are alternative topologies with similar frequencies (ICA close to 0) and even alternative topologies with higher frequency (ICA < 0). The Densitree plot summarized substantial diversity in most of the relationships among the subclades inferred on individual gene trees (Figure 5A). The multidimensional scaling (MDS) analysis showed a diffuse arrangement of gene trees, which indicates there is no dominant cluster of topologies (Figure 5B).

Network analysis

Split networks recovered all clades and subclades obtained in the concatenation and coalescence analyses (Figure 6).

TABLE 2 Characteristics of the exon and supercontigs datasets obtained with the Angiosperms353 bait kit.

Characteristics	Exon dataset without exclusion	Supercontig dataset without exclusion	Exons- with- exclusions dataset (EWE)	Supercontigs- with- exclusions dataset (SCWE)
Total loci analyzed	302	317	258	267
Number of sequences per loci, minimum-maximum of each dataset	29-65	29-65	57-65	60-65
Total sites after trim	212,329	850,115	202,107	785,373
GC percentage	41.83	31.99	41.74	32.03
Informative site percentage	16.11	22.4	16.01	22.51
Indel percentage	3.81	10.68	3.95	10.87

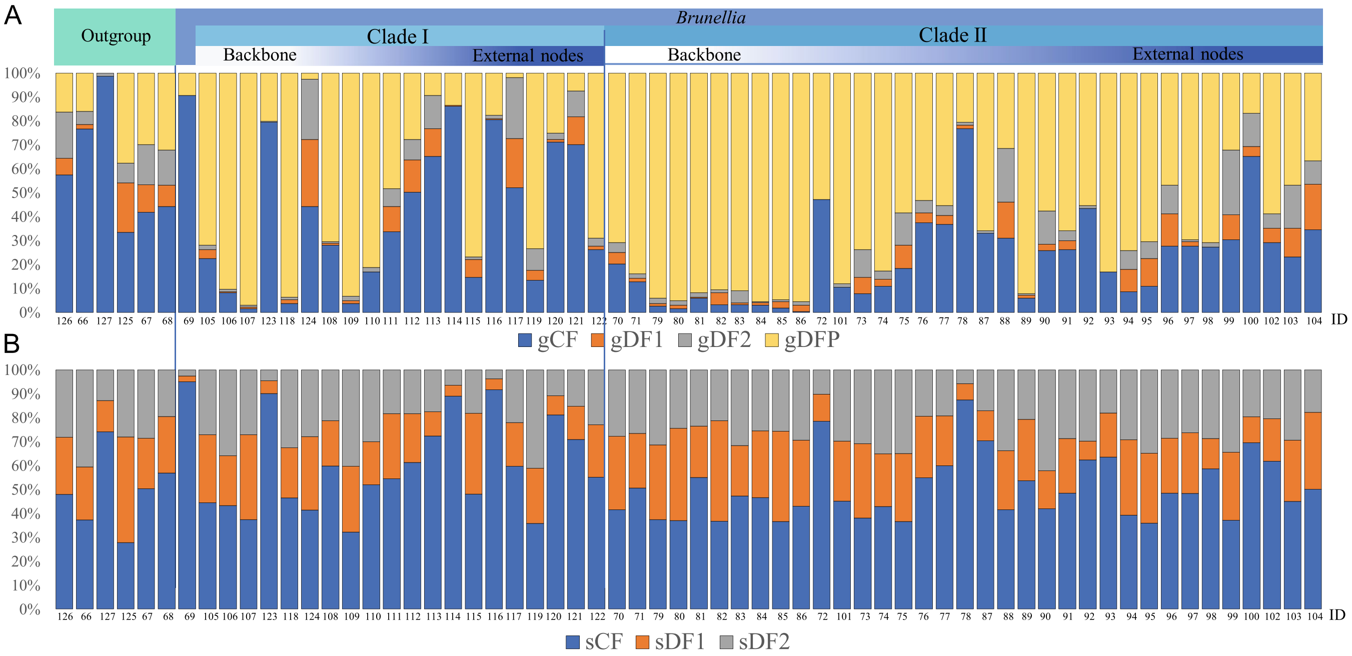

FIGURE 3 Bar-plots showing concordance of genes (gCF, A) and sites (sCF, B), and discordance gene gDF1, gDF2, gDFP) and sites (sCF, sDF1, sDF2) on supercontig dataset of Brunelliaceae with locus exclusion, inferred using maximum likelihood analysis of concatenated data. Numbers of branches (ID) are according to Appendix S10. The ID numbers were arranged by clade and by their position in the tree (backbone or terminal nodes).

The two main clusters (clades I and II) were clearly defined, but high conflict was found among smaller clusters (subclades), and particularly within some of them (subclades Ib, Ic, IIc, and IId).

Hybridization and ILS detection

The HyDe test suggested that hybridization played an important role in the speciation of *Brunellia*. In this test, 87,780 triplets (P1, P2, and a hybrid) were analyzed, but only 4105 (Appendix S14) had considerable levels of hybridization (Bonferroni corrected *P*-value of 0.05/ $87,780 = 5.69 \times 10^{-7}$, Z-score >4.86). The distribution of the γ values, which represent the contribution of the parental genomes, indicated significant levels of hybridization (Appendix S15). Most of the triplets were in the Ib, IIb, IIc, and IId subclades, where evident conflicts were also detected in the network analysis. The phylogenetic terminals with the highest frequency of inferred hybridization are in subclade IIb (*B. bullata* Cuatrec., *B. cayambensis* Cuatrec., and *B. gracilis* C.I. Orozco), and subclade IId (*B. almaguerensis* Cuatrec., *B. boqueronensis* Cuatrec., *B. ecuadoriensis* Cuatrec., and *B. pitayensis* Cuatrec.). *Brunellia coroicoana, B. hygrothermica*, and *B. rhoides* were the three species with the lowest hybridization events in their ancestry (Figure 7, Appendix S14).

Incomplete lineage sorting was inferred for most branches of the SCWE dataset (where the H_0 was accepted). Those events occurred most frequently in branches that led to species with low probability of hybrid origin (Figure 7). In clade I, ILS was mainly detected in the external branches,

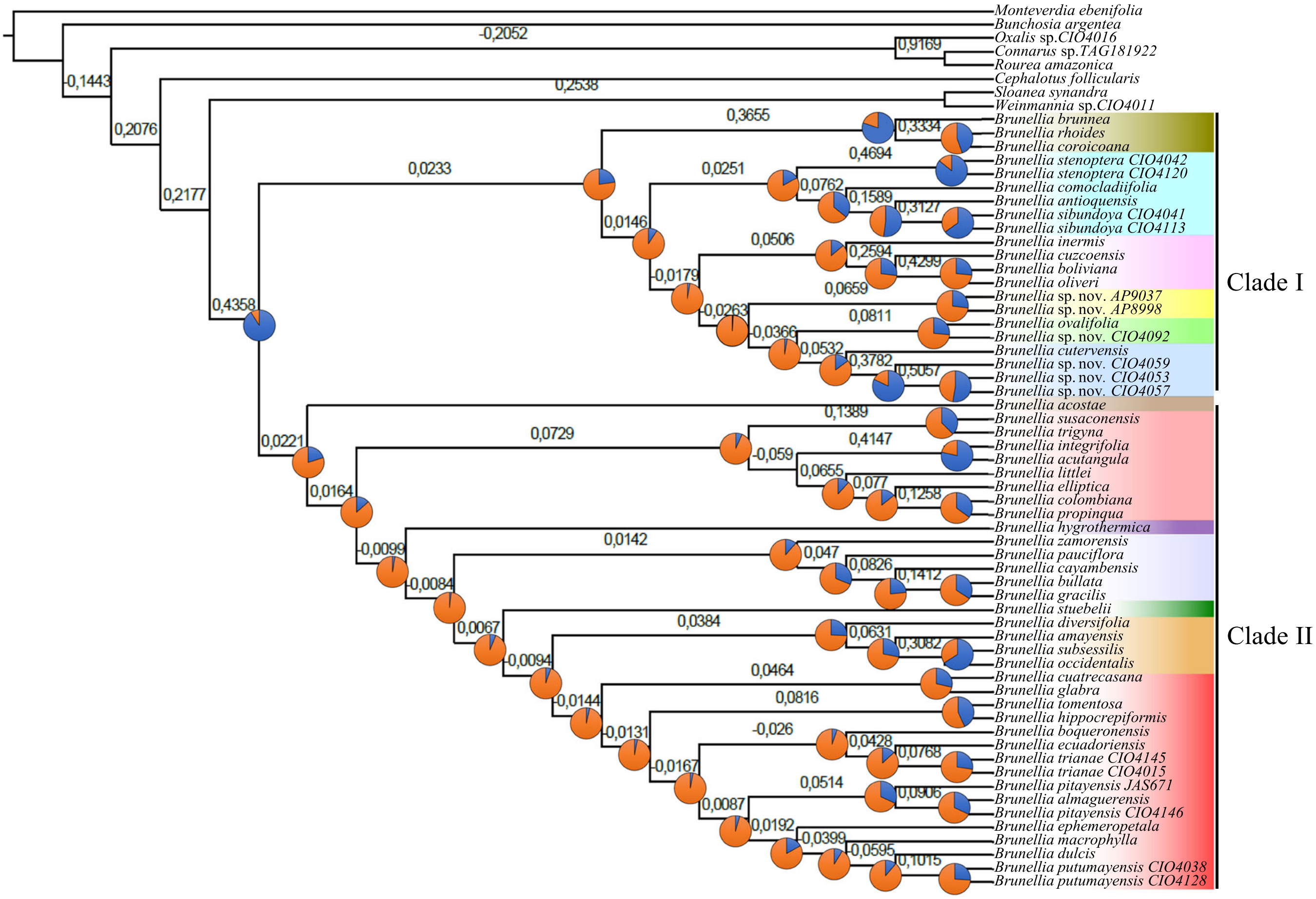

FIGURE 4 ASTRAL-based species tree of Brunelliaceae reconstructed from supercontigs dataset with locus exclusion showing gene tree discordance. Pie charts show the proportion of concordant (blue) and discordant genes (orange). Numbers above branches indicate internode certainty of all (ICA) values.

whereas in clade II they were found in some external branches and in some of the backbone (Figure 7).

Ancestral character-state reconstruction

All examined traits are inferred to be homoplastic. Therefore, none of the traits were associated with any of the two main clades (Figure 8). All characters show a significantly better fit using a one-parameter model (p 0.001), except for leaf complexity, which has a better fit using an asymmetrical 2-parameter model (forward rate: 0.03061, backward rate: 0.106531). Leaf complexity changed from an ancestral state of compound leaves to simple leaves six independent times (Figure 8A). However, most of the simple (unifoliolate)-leaved species belonged to subclade IId, where only *B. ecuadoriensis* has reverted to compound leaves. The complexity of proximal and distal branches of inflorescences has been reduced five times independently. In clade I, most species had thyrsoid inflorescences with higher complexity in proximal and distal branches, and just a few developed less complex inflorescences (Figure 8B). In clade II, it is evident that there was a reduction of inflorescence complexity. Most species develop thyrsoid inflorescences with higher complexity only at the proximal branches, and few species have monothyrsoid inflorescences (Figure 8B). The fertile portion of the inflorescence did not have a clear pattern in terms of the evolution of its character states (Figure 8C).

Calyx merosity was highly homoplastic in clade I; in clade II, most of the species had five calyx lobes, but this character changed multiple times to a lesser or higher number of lobes (Figure 8D). The ancestral condition in the family is inferred to have the number of carpels equal to the calyx lobes (Figure 8E). This condition changed multiple times to a condition with fewer carpels than calyx lobes, mostly in clade II. Carpel number is highly homoplastic within *Brunellia* and does not show unambiguous patterns of evolution (Figure 8F). The shape of the endocarp is inferred to be highly homoplastic (Figure 8G). The U-shape endocarp is reconstructed as the ancestral state in the family. In clade I, the endocarp is inferred to have changed to urceolate and later to navicular, whereas in clade II, the U-shape changed to navicular in subclade IIa and to urceolate in the ancestor of subclades IIb–IId. Therefore, the navicular shape arose from the U-shape and the urceolate shape independently in subclades IIa and IId.

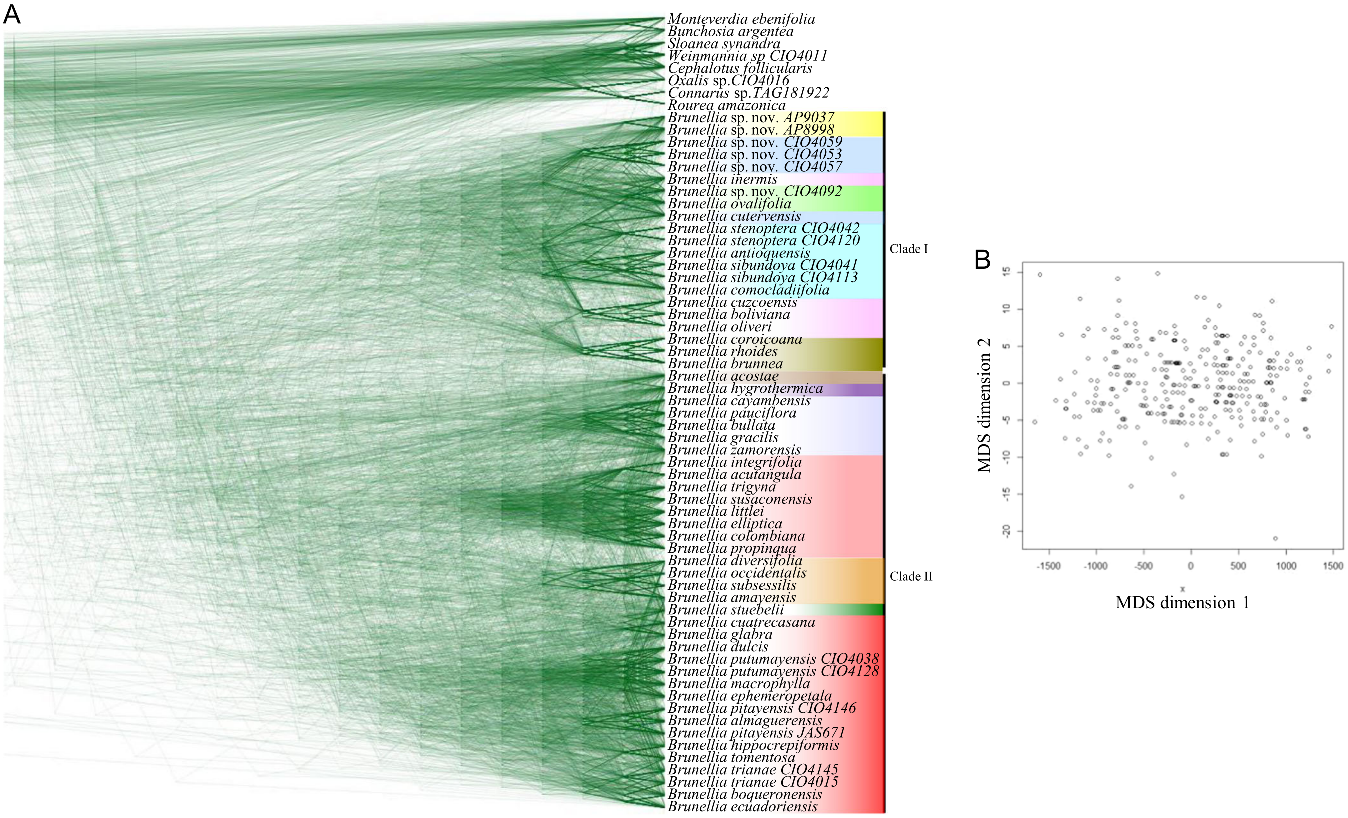

FIGURE 5 Gene trees discordance of Brunelliaceae based on supercontig data set with exclusion. (A) DensiTree plot. (B) Multidimensional scaling plot of Robinson Foulds distances showing the high dispersion of gene trees (data points in the figure), indicating the lack of a unique cluster or topology.

DISCUSSION

Assessment of noise and phylogenetic signal

High-throughput sequencing has emerged as a tool for a better understanding of evolutionary history, recovering more resolved phylogenies with higher support (e.g., for birds, Prum et al., 2015; *Salvia* [Lamiaceae], Fragoso-Martínez et al., 2017; *Nepenthes* [Nepenthaceae], Murphy et al., 2020; see also Rokas et al., 2003). With the amount of information generated by NGS, filtering procedures can help obtain clearer results. Also, it is highly advantageous to have a reference genome in order to accurately detect gene duplication, hybridization, introgression, ILS, saturation, and long-branch attraction, which could add noise and bias in phylogenetic analyses (Straub et al., 2014; Smith et al., 2015; Herrando-Moraira et al., 2018). All of the aforementioned phenomena can mislead phylogenetic reconstruction (Straub et al., 2014; García et al., 2017; Nikolov et al., 2019) and should be taken into account to allow a clearer interpretation of evolutionary history. We used a targeting sequencing technique (Hyb-Seq) here to obtain single-copy nuclear loci that could be phylogenetically informative for the reconstruction of the relationships among the species of *Brunellia*. Here, we present the first molecular phylogeny for this group, assess probable origins of the incongruences among gene trees, and re-examine morphological characters traditionally used in the taxonomy of *Brunellia*.

We evaluated loci recovered with the Angiosperms353 toolkit (Johnson et al., 2019) to detect possible events that could confound the phylogenetic signal. We found 27 loci with paralogs, but only nine were paralogous within *Brunellia*, representing 2.57% of the assembled loci. Paralogs can mislead phylogenetic inference in concatenated analyses, but may not always affect coalescent analyses (e.g., Maddison 1997, Du et al., 2019 [Preprint]; Soto Gomez et al., 2019). For this reason, we decided to exclude ingroup paralogs from all further analyses. We found strong discordance among loci in the phylogenetic analyses of the exon and supercontigs datasets (Appendix S10A, C, E, F). Therefore, we applied ten approaches to filtering out loci that were too noisy or that may lead to biased phylogenetic inference. All loci that had such biases, including those putatively with long branch attraction and excessive sequence saturation, or predicted paralogy, were removed.

When the exon and supercontig datasets are compared, the latter contains almost four times the number of nucleotides of the former (Table 2), and have a slight increase in the gene and site concordance of the branches in the analyses (gCF, sCF, and QS). Exons have a lower rate of DNA sequence evolution and therefore have somewhat lower utility for inferring relationships among very recent or closely related taxa, as reflected in short branches in the

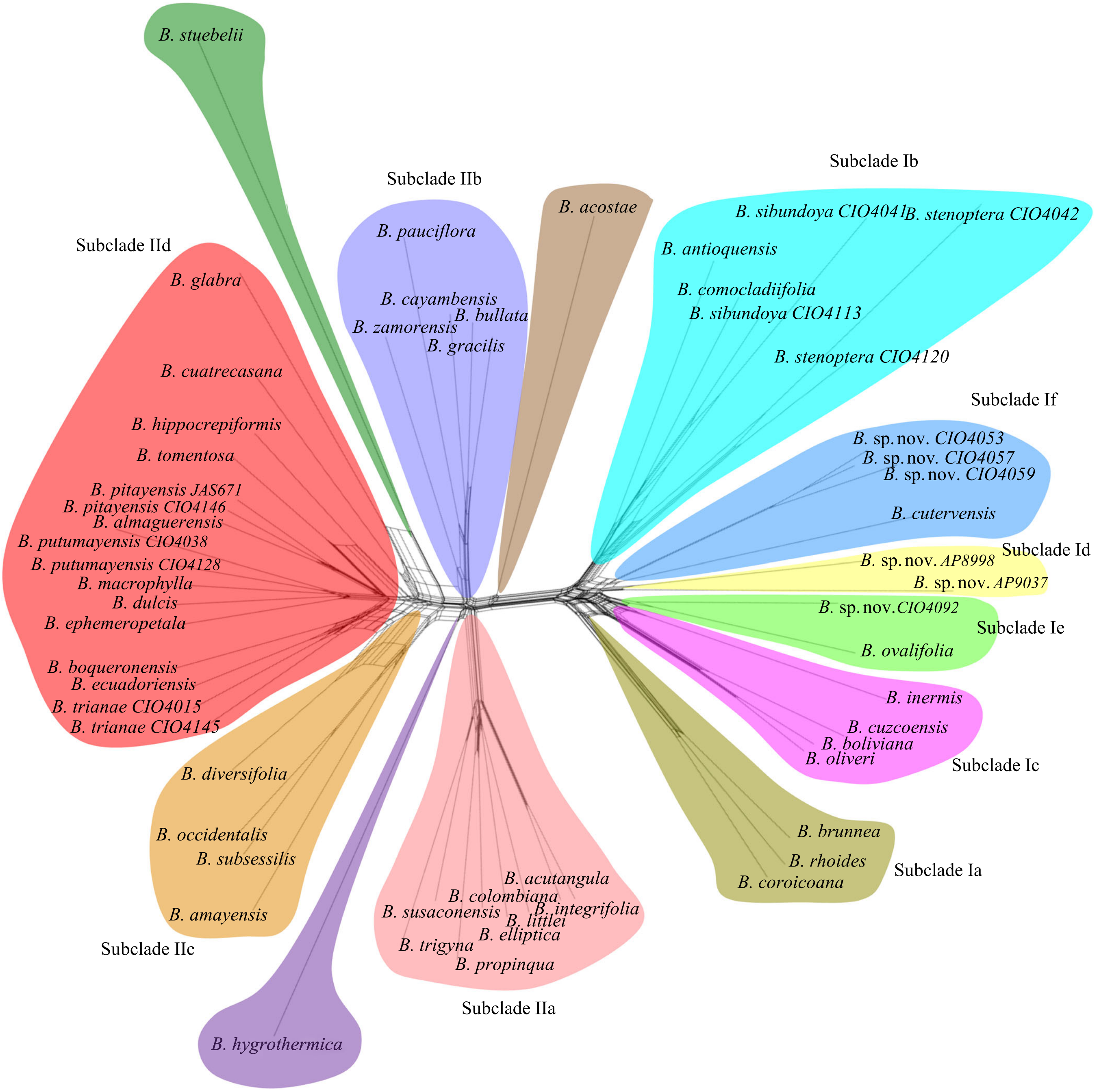

FIGURE 6 Split network of Brunelliaceae estimated with the MedianNetwork method using the supercontigs with exclusion dataset. Same clades and subclades from the phylogenetic analysis are recovered as clusters in the network.

phylogenies. The use of noncoding flanking regions, which generally have a higher rate of evolution, could help to resolve inconsistencies caused by lack of variation in exonic regions. The gene trees inferred from the EWE dataset had lower than average UFBoot support than those from the SCWE dataset. These results indicate that the topologies of the EWE dataset were less consistent within each gene than in the other dataset.

The ML and ASTRAL analyses for the SCWE dataset led to recovery of two main clades (Appendix S10G and Figure 1, respectively): one (I) with six subclades and the other (II) with four subclades and three isolated species. The relationships among the subclades differed among the two methods; however, the relationships among species were consistent in the majority of trees. Trees for other datasets (exons and supercontigs without exclusions and EWE datasets) also recovered the clades and subclades, but relationships among species and among subclades were variable (Appendix S10A–F). Analyses with loci filtered to remove noisy or potentially biased loci can recover more

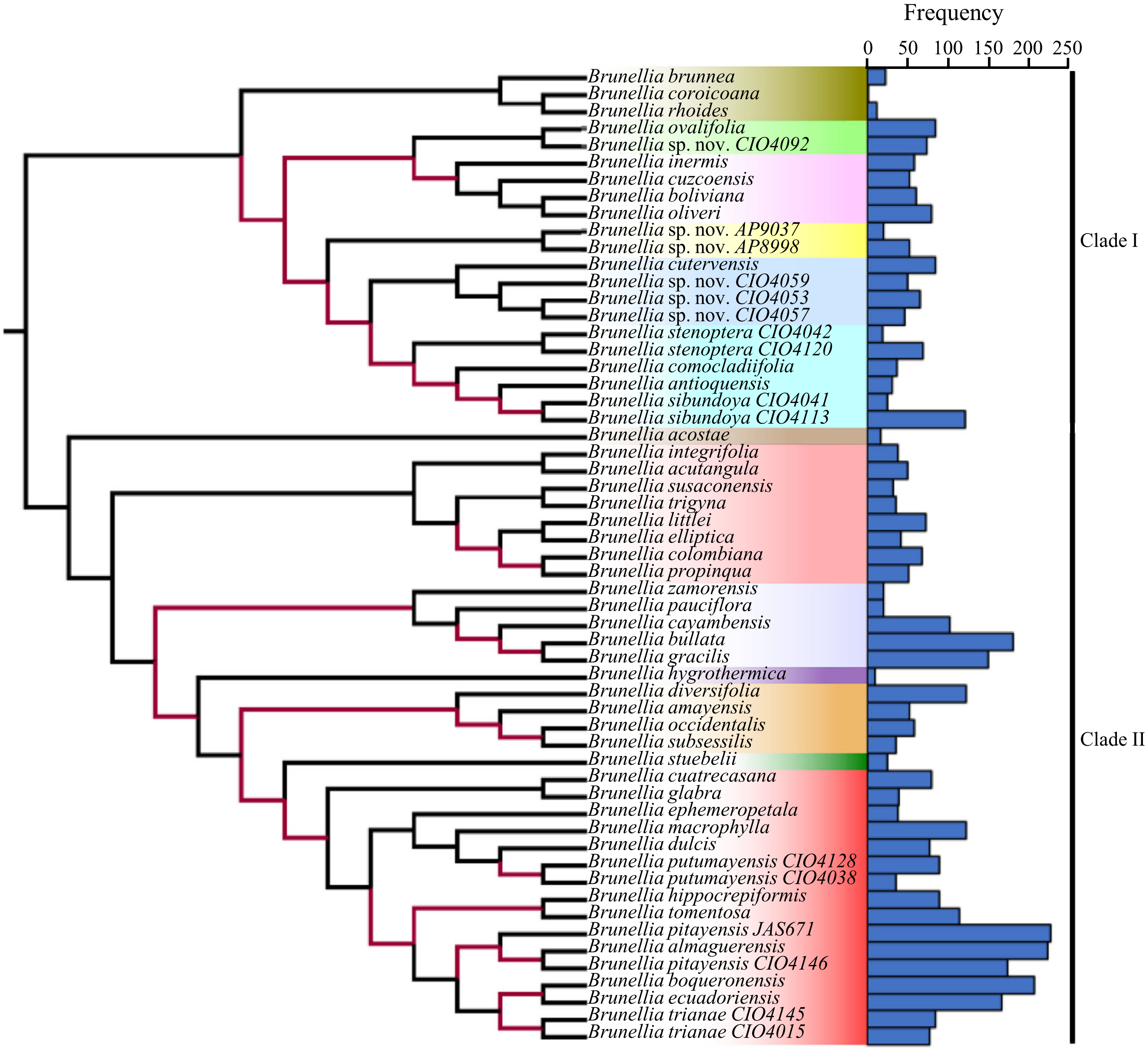

FIGURE 7 ML tree of Brunelliaceae based on the supercontigs with exclusion dataset showing ILS and hybridization. Red branches are those where the hypothesis of ILS was rejected. The bar plot shows for each terminal the frequency where each taxon is predicted to be a hybrid in the hybridization HyDe test (Z-score >4879).

stable phylogenies, with conflicts only in the most recalcitrant clades (e.g., Nikolov et al., 2019), whose recalcitrance can be attributed to complex speciation processes. However, after performing loci filtering to reduce noise and phylogenetic bias, we did not recover consistent relationships among subclades or improved gene and site concordance in *Brunellia*. These discrepancies have been found in other studies, showing that increasing the number of genes does not necessarily increase support and topological congruence (e.g., Bogarín et al., 2018). This may reflect variation in the rate of nucleotide substitution of the genes (Bagley et al., 2020) and biological processes that are unique to the organisms studied (e.g., Chen et al., 2009).

Gene-tree discordance

In general, all branches within *Brunellia* are very short in the ML analysis (Appendix S12) and have few genes and sites that support the main topology (Figure 2), which is consistent with gene concordance factor (gCF) values mainly <30%, sites concordance factor (sCF) values mainly <50% (Appendix S10G), most nodes with QS < 50% (Figure 1). Nonetheless, branch support as measured by the bootstrap analysis (UFBoot) and local posterior probabilities (LPP) were high in most branches (74% of branches have UFBoot = 100%, 78% of the branches have LPP > 0.9); however, these statistical measures are sensitive in genomic datasets to biases related to

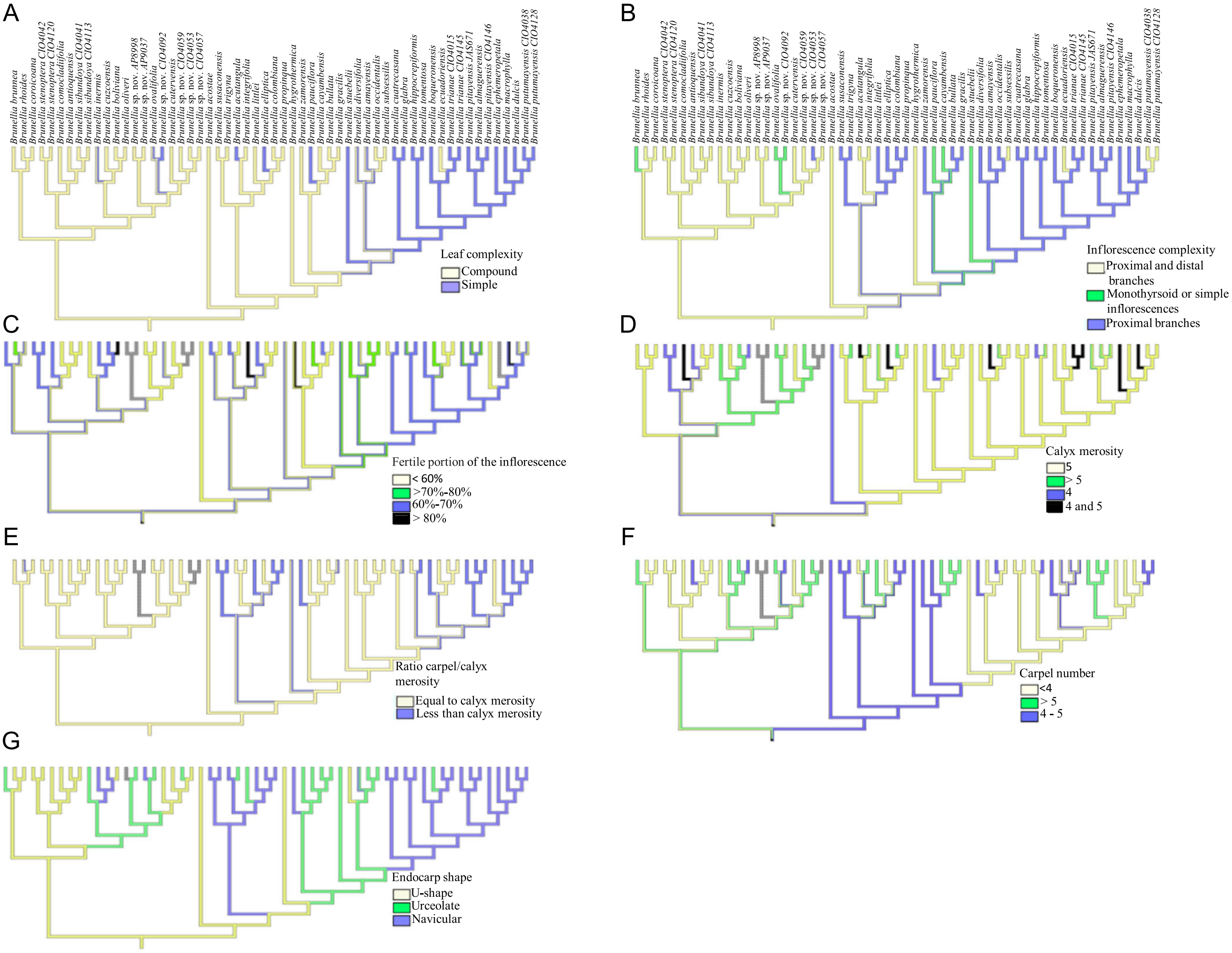

FIGURE 8 Ancestral character state reconstruction for seven morphological traits of *Brunellia*. (A) Leaf complexity, (B) inflorescence complexity, (C) fertile portion of the inflorescence, (D) calyx merosity, (E) ratio carpel/calyx merosity, (F) carpel number, (G) endocarp shape. Gray branches indicate that the character was not evaluated in that taxon.

data quantity (as support values are dependent on the amount of genes/data analyzed, e.g., Rokas et al., 2003; Phillips et al., 2004; Soltis et al., 2004). There was also a low proportion of genes supporting the two main alternative topologies (gDF1 and gDF2) and most of the genes supporting any different one (gDFP > 70.77% for the SCWE dataset). Additionally, there was a relatively high proportion of sites supporting the branches of the main topology (sCF > 40% for the SCWE dataset) versus those that supported alternative topologies (sDF mainly ~20–30%). In the backbone of each main clade, the relationships presented a higher level of gene discordance and lower ICA values than the nodes near the tips (Figure 4). This reflects the presence of alternative topologies with similar frequency or even with higher frequency. This type of discordance can be caused by ILS, especially when there is rapid and recent radiation (Whitfield and Lockhart, 2007; Bagley et al., 2020), which has also been found in other groups (Jabaily et al., 2018; Bagley et al., 2020).

Historical processes

The network analysis and the Densitree plot (Figures 5 and 6), reveal high discordance in our data, implying conflicts among gene trees, which could be caused by duplication, reticulate evolution, and ILS (Maddison, 1997; Wendel and Doyle, 1998; Som, 2014; Smith et al., 2015; Bogarín et al., 2018). However, the Angiosperm353 kits targets single-copy genes, which should be less subject to paralogy issues (e.g., Johnson et al., 2016; Baker et al., 2021), paralogs found using HybPiper and tree-to-tree distances were removed from the analyses, and Phyparts did not detect gene duplication in our datasets. However, our other analysis detected strong evidence for hybridization and ILS (Figure 7; Appendix S14). ILS has been associated with rapid radiation events (e.g., Mao et al., 2019), which could be the case in *Brunellia* based on the very short branches detected here in the ML analysis (Appendix S12). All of these events could generate conflict among reconstructed phylogenies inferred from diverse datasets (Mao et al., 2019). Hybridization is one of the major forces in plant evolution contributing to speciation (e.g., Grant, 1981; Rieseberg, 1997; Payseur and Rieseberg, 2016) and is reported to co-occur with recent radiation and rapid adaptation to new environments (e.g., Grant, 1981; Rieseberg, 1997; Mallet, 2007; Marques et al., 2019, Morales-Briones et al., 2021). Here, we found that only three species (*B. acostae*, *B. hygrothermica*, and *B. stuebelii*) appear as probable parents of hybrids in 47.32% of the hybridization tests (Appendix S14). Conflict among gene trees can also reflect the presence of alternative alleles that remain in the different lineages through successive speciation events, known as ILS (e.g., Pamilo and Nei, 1988; Maddison, 1997; Whitfield and Lockhart, 2007; Oliver, 2013; Jabaily et al., 2018). ILS was observed in around 50% of the branches in the phylogeny of *Brunellia* (Figure 7), which likely explains the lack of a consistent topology among the gene trees evaluated here.

Phylogeny and morphological character evaluation

We recovered two clades in Oxalidales, one formed by Oxalidaceae + Connaraceae and the other by Cephalotaceae + (Brunelliaceae + Cunoniaceae + Elaeocarpaceae). This is consistent with Soltis et al. (2011), who used 17 genes from the three genomes (nuclear genome, plastome, and mitogenome), and by Li et al. (2021), who used 76 plastid protein-coding genes. In the latter clade (which includes Brunelliaceae), our data indicated that Cephalotaceae was sister to the other families, in agreement with Wurdack and Davis (2009), Qiu et al. (2010), Bradford and Barnes (2001), Sun et al. (2016), and Li et al. (2021). However, this disagrees with the relationships found by Soltis et al. (2011) where Brunelliaceae was sister to Elaeocarpaceae (Cephalotaceae + Cunoniaceae). Most of our phylogenetic analyses recovered Brunelliaceae sister to Cunoniaceae + Elaeocarpaceae. Three of the analyses with the exon dataset recovered Brunelliaceae sister to Elaeocarpaceae, as previously found by Pillon et al. (2021), who also used the Angiosperms353 toolkit. However, the relationships from Pillon et al. (2021) may be an artifact of taxon sampling, as they included only one Brunelliaceae accession and 37 taxa from Cunoniaceae.

Our analyses recovered Brunelliaceae as monophyletic with high support. This result was previously found in a study based on morphological characters (Orozco, 2001) and plastid markers (Bradford and Barnes, 2001). Also, all our analyses recovered two large clades in Brunelliaceae, 10 subclades and three isolated species. Infrageneric classifications of *Brunellia* (sections and subsections) based on morphological characters (Cuatrecasas, 1970, 1985; Orozco, 2001) partly conflict with our results. Cuatrecasas (1970, 1985) used leaf complexity for delimiting sections, i.e., section *Brunellia* was distinguished by the presence of compound leaves, whereas section *Simplicifolia* was characterized by the presence of simple leaves. Neither section is monophyletic, and the two main clades (I and II) recovered in our phylogenetic analyses contain members of both sections. The reduction of leaflets and floral parts was reconstructed as arising multiple times independently in *Brunellia*, which is why the traditional classification does not reflect the phylogenetic relationships. These reductions were probably influenced by the colonization of the new environments formed during the final uplift of the Andes during the Quaternary, as has happened in other groups of plants (Gómez-Gutiérrez et al., 2017).

Our results also do not support the infrageneric classification of Orozco (2001), who delimited five sections within *Brunellia*, although subclade IId contains most (but not all) of the species classified in sect. *Simplicifolia* by Cuatrecasas (1985) and Orozco (2001) on the basis of having simple leaves and floral reduction. Subclade IIa contains all species included in sect. *Brunellia* subsect. *Colombianae* of Cuatrecasas (1985) and sect. *Simplicifoliae* subsect. *Propinquae* from Orozco (2001), with the addition of *B. acutangula* Bonpl., *B. integrifolia* Szyszyl., and *B. susaconensis* (Cuatrec.) C.I. Orozco. Most of the species of sect. *Brunellia* were recovered in clade IA, except for several undescribed species and *B. ovalifolia* Bonpl. Other sections and subsections are not recovered in the molecular analyses. The main clades of *Brunellia* cannot be defined with any of the morphological characters studied here, as they are defined on the basis of highly homoplastic traits (Figure 8; Appendix S16).

Despite unclear morphological diagnosability in *Brunellia*, geographic patterns can be aligned with our results. In general, clade I is more broadly distributed than clade II. *Brunellia* species of clade I grow from Colombia to Peru and Bolivia, with one species, *B. comocladiifolia* Bonpl., widely distributed from Costa Rica to Ecuador, Venezuela, and the Greater Antilles (Cuatrecasas, 1970, 1985; Orozco, 2001). Species of clade II are mainly distributed in Colombia, Ecuador and Venezuela, but *B. dulcis* J.F. Macbr. is endemic to Peru, and *B. hygrothermica* reaches Panama.

CONCLUSIONS

We found evidence to explain the topological variability and the low branch support in *Brunellia*, consistent with the topological incongruence between the coalescent and concatenated methods and the high discordance between the gene trees and the species trees. This could be due to various factors, including a lack of phylogenetic signal, hybridization, and high ILS associated with this rapid radiation, which is consistent with results found in other groups (Smith et al., 2015; Jabaily et al., 2018; Widhelm et al., 2019). The precise causes of discordance remain a fundamentally difficult set of phenomena to disentangle (Morales-Briones et al., 2021).

AUTHOR CONTRIBUTIONS

J.M.A., J.V.D., C.I.O., C.P.O., and K.M.N. conceived and planned the experiments. J.M.A., C.I.O., and C.P.O. collected the samples in the field. J.V.D. and K.M.N. prepared laboratory samples. J.V.D. performed bioinformatic raw data processing. J.M.A. analyzed the data; J.M.A., J.V.D., and K.M.N. prepared figures and tables; all authors authored or reviewed drafts of the paper, and approved the final manuscript.

ACKNOWLEDGMENTS

This project was funded by the Departamento Administrativo de Ciencia, Tecnología e Innovación (Colciencias), project Colciencias-1101658, granted to C.I.O., J.M.A., and C.A.P. Also, we received support from División de Investigación y extensión (DIEB) of the Universidad Nacional de Colombia, Bogotá, project DIEB-15858, granted to C.I.O. and J.M.A. J.V.D. was supported by a Fulbright-Colciencias scholarship. Lab work was supported by start-up funds from Southern Illinois University, NSF DUE-1564969 and DBI-2002400 to K.M.N. The authors thank the curators of CAUP, COL, CUVC, F, HOXA, HUA, LBP, MEDEL, MO, NY, QCA, US, USM, and UTMC for letting us study their collections, to the computer engineers William Conrad Phillip and Emilia Zeledón Lostalo for their help in the development of scripts and assistance with the programs used in this research, and to Silvia Bustamante and her team at the Instituto de Biotecnología, Universidad Nacional de Colombia (IBUN), for their assistance in the first stage of the project. Also, we thank Segundo Leyva (HAO), Andrés Felipe Bohórquez (COL), Andrés Chitivo (COL), and Carlos Velasco (COL) for their help with fieldwork. The authors thank the anonymous reviewers and editors for their comments and contributions to the improvement of the manuscript.

DATA AVAILABILITY STATEMENT

Vouchers were deposited in public herbaria (Appendix S1). Reads, assemblies, contigs captured and obtained using HybPiper, as well as all matrices and trees used in this study are available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.3tx95x6jd

ORCID

José Murillo-A. https://orcid.org/0000-0002-3703-4250
Janice Valencia-D. https://orcid.org/0000-0003-2508-9825
Clara I. Orozco https://orcid.org/0000-0001-5639-2558
Carlos Parra-O. https://orcid.org/0000-0002-9807-4619
Kurt M. Neubig https://orcid.org/0000-0002-7113-0449

REFERENCES

Bagley, J. C., S. Uribe-Convers, M. M. Carlsen, and N. Muchhala. 2020. Utility of targeted sequence capture for phylogenomics in rapid, recent angiosperm radiations: Neotropical *Burmeistera* bellflowers as a case study. *Molecular Phylogenetics and Evolution* 152: 106769.

Baker, W. J., S. Dodsworth, F. Forest, S. W. Graham, M. G. Johnson, A. McDonnell, L. Pokorny, et al. 2021. Exploring Angiosperms353: An open, community toolkit for collaborative phylogenomic research on flowering plants. *American Journal of Botany* 108: 1059–1065.

Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. Lesin, et al. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* 19: 455–477.

Blischak, P. D., J. Chifman, A. D. Wolfe, and L. S. Kubatko. 2018. HyDe: A Python package for genome-scale hybridization detection. *Systematic Biology* 67: 821–829.

Bogarín, D., O. A. Pérez-Escobar, D. Groenenberg, S. D. Holland, A. P. Karremans, E. M. Lemmon, A. R. Lemmon, et al. 2018. Anchored hybrid enrichment generated nuclear, plastid and mitochondrial markers resolve the *Lepanthes horrida* (Orchidaceae: Pleurothallidinae) species complex. *Molecular Phylogenetics and Evolution* 129: 27–47.

Bogdanowicz, D., K. Giaro, and B. Wróbel. 2012. TreeCmp: Comparison of trees in polynomial time. *Evolutionary Bioinformatics* 8: 475–487.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114–2120.

Bouckaert, R. R. 2010. DensiTree: Making sense of sets of phylogenetic trees. *Bioinformatics* 26: 1372–1373.

Capella-Gutiérrez, S., J. M. Silla-Martínez, and T. Gabaldón. 2009. TrimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25: 1972–1973.

Chen, J. Q., Y. Wu, H. Yang, J. Bergelson, M. Kreitman, and D. Tian. 2009. Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. *Molecular Biology and Evolution* 26: 1523–1531.

Cuatrecasas, J. 1970. Brunelliaceae. Flora Neotropica monograph 2. Hafner, New York, New York, USA.

Cuatrecasas J. 1985. Brunelliaceae. Flora Neotropica monograph 2 (suppl.). Hafner, New York, New York, USA.

Doyle, J., and J. L. Doyle. 1987. Genomic plant DNA preparation from fresh tissue-CTAB method. *Phytochemical Bulletin* 19: 11–15.

Du, P., M. W. Hahn, and L. Nakhleh. 2019. Species tree inference under the multispecies coalescent on data with paralogs is accurate. *BioRxiv*. Website: https://doi.org/10.1101/498378 [preprint].

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* 27: 401–410.

Fragoso-Martínez, I., G. A. Salazar, M. Martínez-Gordillo, S. Magallón, L. Sánchez-Reyes, E. M. Lemmon, A. R. Lemmon, et al. 2017. A pilot study applying the plant Anchored Hybrid Enrichment method to New World stages. (*Salvia* subgenus *Calosphace*; Lamiaceae). *Molecular Phylogenetics and Evolution* 117: 124–134.

Gadagkar, S., M. S. Rosenberg, and S. Kumar. 2005. Inferring species phylogenies from multiple genes: Concatenated sequence tree versus consensus gene tree. *Journal of Experimental Zoology, Part B, Molecular and Developmental Evolution* 304: 64–74.

García, N., R. A. Folk, A. W. Meerow, S. Chamala, M. A. Gitzendanner, R. Souza de Oliveira, D. E. Soltis, et al. 2017. Deep reticulation and incomplete lineage sorting obscure the diploid phylogeny of rain-lilies and allies. (Amaryllidaceae tribe Hippeastreae). *Molecular Phylogenetics and Evolution* 111: 231–247.

Gómez-Gutiérrez, M. C., R. T. Pennington, L. E. Neaves, R. I. Milne, S. Madriñán, and J. R. Richardson. 2017. Genetic diversity in the Andes: Variation within and between the South American species of *Oreobolus* R. Br. (Cyperaceae). *Alpine Botany* 127: 155–170.

Grant, V. 1981. Plant speciation. Columbia University Press, New York, New York, USA.

Heather, J. M., and B. Chain. 2016. The sequence of sequencers: The history of sequencing DNA. *Genomics* 107: 1–8.

Hendy, M. D., and D. Penny. 1989. A framework for the quantitative study of evolutionary trees. *Systematic Biology* 38: 297–309.

Herrando-Moraira, S., J. A. Calleja, P. Carnicero, K. Fujikawa, M. Galbany-Casals, N. Garcia-Jacas, I. Hyoung-Tak, et al. 2018. Exploring data processing strategies in NGS target enrichment to disentangle radiations in the tribe *Cardueae* (Compositae). *Molecular Phylogenetics and Evolution* 128: 69–87.

Hoang, D. T., O. Chernomor, A. von Haeseler, B. Q. Minh, and L. S. Vinh. 2018. UFBoot2: Improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution* 35: 518–522.

Huson, D. H., and D. Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23: 254–267.

Huson, D. H., T., Klöpper, P. J. Lockhart, and M. A. Steel. 2005. Reconstruction of reticulate networks from gene trees. *In* S. Miyano, J. Mesirov, S. Kasif, S. Istrail, P. Pevzner, and M. Waterman [eds.], Research in computational molecular biology, 233–249. Springer-Verlag, Berlin and Heidelberg, Germany.

Jabaily, R. S., K. A. Shepherd, P. S. Michener, C. J. Bush, R. Rivero, A. G. Gardner, and E. B. Sessa. 2018. Employing hypothesis testing and data from multiple genomic compartments to resolve recalcitrant backbone nodes in *Goodenia* sl (Goodeniaceae). *Molecular Phylogenetics and Evolution* 127: 502–512.

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

Johnson, M. G., L. Pokorny, S. Dodsworth, L. R. Botigué, R. S. Cowan, A. Devault, W. L. Eiserhardt, et al. 2019. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant esigned using k-medoids clustering. *Systematic Biology* 68: 594–606.

Joly, S. 2012. JML: Testing hybridization from species trees. *Molecular Ecology Resources* 12: 179–184.

Kalyaanamoorthy, S., B. Q. Minh, T. K. F. Wong, A. von Haeseler, and L. S. Jermiin. 2017. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Natural Methods* 14: 587–589.

Kanzi, A. M., J. E. San, B. Chimukangara, E. Wilkinson, M. Fish, V. Ramsuran, and T. de Oliveira. 2020. Next generation sequencing and bioinformatics analysis of family genetic inheritance. *Frontiers in Genetics* 11: 544162.

Katoh, K., and D. M. Standley. 2013. MAFFT Multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* 30: 772–780.

Kück, P., and G. Longo. 2014. FASconCAT-G: Extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. *Frontiers in Zoology* 11: 81.

Kumar, S., A. Filipski, F. Battistuzzi, S. Kosakovsky, and K. Tamura. 2012. Statistics and truth in phylogenomics. *Molecular Biology Evolution* 29: 457–472.

Lanfear, R. 2018. Calculating and interpreting gene- and site-concordance factors in phylogenomics. Molecular evolution and phylogenetics blog. Website: http://www.robertlanfear.com/blog/files/concordance_factors.html [accessed 2 March 2022].

Lewis, P. O. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology* 50: 913–925.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, et al., 1000 Genome project data processing subgroup. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25: 2078–2079.

Li, X., Y. Zhao, X. Tu, C. Li, Y. Zhu, H. Zhong, Z. Liu, et al. 2021. Comparative analysis of plastomes in Oxalidaceae: Phylogenetic relationships and potential molecular markers. *Plant Diversity* 43: 281–291.

Maddison, W. P. 1997. Gene trees in species trees. *Systematic Biology* 46: 523–536.

Maddison, W. P., and D. R. Maddison. 2018. Mesquite: A modular system for evolutionary analysis, version 3.51. Website: http://www.mesquiteproject.org

Mallet, J. 2007. Hybrid speciation. *Nature* 466: 279–283.

Mao, K., M. Ruhsam, Y. Ma, S. W. Graham, J. Liu, P. Thomas, R. I. Milne, and P. M. Hollingsworth. 2019. A transcriptome-based resolution for a key taxonomic controversy in Cupressaceae. *Annals of Botany* 123: 153–167.

Marques, D. A., J. I. Meier, and O. Seehausen. 2019. A combinatorial view on speciation and adaptive radiation. *Trends in Ecology and Evolution* 34: 531–544.

McKain, M. R., M. G. Johnson, S. Uribe-Convers, D. Eaton, and Y. Yang. 2018. Practical considerations for plant phylogenomics. *Applications in Plant Sciences* 6: e1038.

Minh, B. Q., M. W. Hahn, and R. Lanfear 2020. New methods to calculate concordance factors for phylogenomic datasets. *Molecular Biology and Evolution* 37: 2727–2733.

Minh, B. Q., H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, and R. Lanfear. 2020. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution* 37: 1530–1534.

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

Morales-Briones, D. F., D. T. Tefarikis, M. J. Moore, S. Smith, S. F. Brockington, A. Timoneda, W. C. Yim, et al. 2021. Disentangling sources of gene tree discordance in phylogenomic data sets: Testing ancient hybridizations in Amaranthaceae. *Systematic Biology* 70: 219–235.

Morey, M., A. Fernández-Marmiesse, D. Castiñeiras, J. M. Fraga, M. L. Couce, and J. A. Cocho. 2013. A glimpse into past, present, and future DNA sequencing. *Molecular Genetics and Metabolism* 110: 3–24.

Murphy, B., F. Forest, T. Barraclough, J. Rosindell, S. Bellot, R. Cowan, M. Golos, et al. 2020. A phylogenomic analysis of *Nepenthes* (Nepenthaceae). *Molecular Phylogenetics and Evolution* 144: 106668.

Neubig, K. M., W. M. Whitten, J. R. Abbott, S. Elliott, D. E. Soltis, and P. Soltis. 2014. Variables affecting DNA preservation in archival plant specimens. *In* W. Applequist and L. Campbell [eds.], DNA banking for the 21st century: Proceedings of the U.S. workshop on DNA banking, January 2013, 81–112. William L. Brown Center, Missouri Botanical Garden, St. Louis, USA.

Nikolov, L. A., P. Shushkov, B. Nevado, X. Gan, I. A. Al-Shehbaz, D. Filatov, C. D. Bailey, et al. 2019. Resolving the backbone of the Brassicaceae phylogeny for investigating trait diversity. *New Phytologist* 222: 1638–1651.

Nowrousian, M. 2010. Next-generation sequencing techniques for eukaryotic microorganisms: Sequencing-based solutions to biological problems. *Eukaryotic Cell* 9: 1300–1310.

Oakley, T. H., M. A. Alexandrou, R. Ngo, M. S. Pankey, C. C. K. Churchill, and K. B. Lopker. 2014. Osiris: Accessible and reproducible phylogenetic and phylogenomic analyses within the Galaxy workflow management system. *BMC Bioinformatics* 15: 230.

Oliver, J. C. 2013. Microevolutionary processes generate phylogenomic discordance at ancient divergences. *Evolution* 67: 1823–1830.

Orozco, C. I. 2001. Evolutionary biology of *Brunellia* Ruíz and Pavón (Brunelliaceae, Oxalidales). Ph.D. dissertation, University of Amsterdam, Amsterdam, Netherlands.

Orozco, C. I., A. Pérez, K. Romoleroux, A. F. Bohorquez, and J. Murillo-A. 2020. Three new species of the Andean genus *Brunellia* (Brunelliaceae) from Colombia and Ecuador. *Phytotaxa* 433: 27–40.

Pagel, M. 1999. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Systematic Biology* 48: 612–622.

Pamilo, P., and M. Nei. 1988. Relationships between gene trees and species trees. *Molecular Biology and Evolution* 5: 568–583.

Payseur, B. A., and L. H. Rieseberg. 2016. A genomic perspective on hybridization and speciation. *Molecular Ecology* 25: 2337–2360.

Phillips, M. J., F. Delsuc, and D. Penny. 2004. Genome-scale phylogeny and the detection of systematic biases. *Molecular Biology and Evolution* 21: 1455–1458.

Pillon, Y., H. C. F. Hopkins, O. Maurin, N. Epitawalage, J. Bradford, Z. S Rogers, W. J. Baker, and F. Forest. 2021. Phylogenomics and biogeography of Cunoniaceae (Oxalidales) with complete generic sampling and taxonomic realignments. *American Journal of Botany* 108: 1181–1200.

Price, M. N., P. S. Dehal, and A. P. Arkin. 2010. FastTree 2 – Approximately maximum-likelihood trees for large alignments. *PLoS One* 5(3): e9490.

Prum, R., J. Berv, A. Dornburg, D. J. Field, J. P. Townsend, E. Moriarty, and A. Lemmon. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature* 526: 569–573.

Qiu, Y. L., L. Li, B. Wang, J. Y. Xue, T. A. Hendry, R. Q. Li, J. W. Brown, et al. 2010. Angiosperm phylogeny inferred from sequences of four mitochondrial genes. *Journal of Systematics and Evolution* 48: 391–425.

R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Website: http://www.r-project.org/index.html [accessed March 2020].

Rieseberg, L. H. 1997. Hybrid origins of plant species. *Annual Review of Ecology and Systematics* 28: 359–389.

Rieseberg, L. H. and J. H. Willis 2007. Plant speciation. *Science* 317: 910–914.

Robinson, D. F., and L. R. Foulds. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences* 53: 131–147.

Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425: 798–804.

Salichos, L., A. Stamatakis, and A. Rokas. 2014. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. *Molecular Biology and Evolution* 31: 1261–1271.

Soto Gomez, M., L. Pokorny, M. B. Kantar, F. Forest, I. J. Leitch, B. Gravendeel, P. Wilkin, et al. 2019. A customized nuclear target enrichment approach for developing a phylogenomic baseline for *Dioscorea* yams (Dioscoreaceae). *Applications in Plant Sciences* 7: e11254.

Straub, S. C. K., M. J. Moore, P. S. Soltis, D. E. Soltis, A. Liston, and T. Livshultz. 2014. Phylogenetic signal detection from an ancient rapid radiation: Effects of noise reduction, long-branch attraction, and model selection in crown clade Apocynaceae. *Molecular Phylogenetics and Evolution* 80: 169–185.

Shen, X. X., C. T. Hittinger, and A. Rokas. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecology and Evolution* 1: 0126.

Smith, S. A., M. J. Moore, J. W. Brown, and Y. Yang. 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evolutionary Biology* 15: 150.

Soltis, D. E., V. A. Albert, V. Savolainen, K. Hilu, Y. L. Qiu, M. W. Chase, J. S. Farris, et al. 2004. Genome-scale data, angiosperm relationships, and 'ending incongruence': A cautionary tale in phylogenetics. *Trends in Plant Science* 9: 477–483.

Soltis, D. E., S. A. Smith, N. Cellinese, K. J. Wurdack, D. C. Tank, S. F. Brockington, N. F. Refulio-Rodriguez, et al. 2011. Angiosperm phylogeny: 17 genes, 640 taxa. *American Journal of Botany* 98: 704–730.

Som, A. 2014. Causes, consequences and solutions of phylogenetic incongruence. *Briefings in Bioinformatics* 16: 536–548.

Struck, T. H. 2014. TreSpEx—Detection of misleading signal in phylogenetic reconstructions based on tree information. *Evolutionary Bioinformatics Online* 10: 51–67.

Sun, M., R. Naeem, J. X. Su, Z. Y. Cao, J. G. Burleigh, P. S. Soltis, D. E. Soltis, et al. 2016. Phylogeny of the Rosidae: A dense taxon sampling analysis. *Journal of Systematics and Evolution* 54: 363–391.

Sun, M., D. E. Soltis, P. S. Soltis, X. Zhu, J. G. Burleigh, and Z. Chen. 2015. Deep phylogenetic incongruence in the angiosperm clade Rosidae. *Molecular Phylogenetics and Evolution* 83: 156–166.

Tange, O. 2018. Gnu Parallel 2018. Website: https://doi.org/10.5281/zenodo.1146014 [accessed March 2020].

Thiers, B. 2020. *Index Herbariorum*: A global directory of public herbaria and associated staff. New York Botanical Garden's Virtual Herbarium. New York Botanical Garden, Bronx, New York, USA. Website: http://sweetgum.nybg.org/science/ih/ [accessed March 2020].

Valencia, J., J. Murillo-A., C. I. Orozco, C. Parra, and K. M. Neubig. 2020. Complete plastid genome sequences of two species of the Neotropical genus *Brunellia* (Brunelliaceae). *PeerJ* 8: e8392. Website: https://doi.org/10.7717/peerj.8392

Veitia, R. 2005. Paralogs in polyploids: One for all and all for one? *Plant Cell* 17: 4–11.

Weitemier, K., S. C. K. Straub, R. C. Cronn, M. Fishbein, R. Schmick, A. McDonnell, and A. Liston. 2014. Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences* 2: 1400042.

Wendel, J. F., and J. J. Doyle. 1998. Phylogenetic incongruence: Window into genome history and molecular volution. *In* D. E. Soltis, P. S. Soltis, and J. J. Doyle [eds.], Molecular systematics of plants II: DNA sequencing, 265–296. Kluwer Academic Publishers, Norwell, Massachusetts, USA.

Whitfield, J. B., and P. J. Lockhart. 2007. Deciphering ancient rapid radiations. *Trends in Ecology and Evolution* 22: 258–265.

Widhelm, T. J., F. Grewe, J. P. Huang, J. A. Mercado-Díaz, B. Goffinet, R. Lücking, B. Moncada, et al. 2019. Multiple historical processes obscure phylogenetic relationships in a taxonomically difficult group (Lobariaceae, Ascomycota). *Scientific Reports* 9: 8968.

Williams, C. 2022. addNewSupercontigs. Python script available from the author, website: https://github.com/PopGen33/addNewSupercontigs [accessed 14 February 2022].

Wolfe, K. H. 2001. Yesterday's polyploids and the mystery of diploidization. *Nature Reviews Genetics* 2: 333–341.

Wurdack, K. J., and C. C. Davis. 2009. Malpighiales phylogenetics: Gaining ground on one of the most recalcitrant clades in the angiosperm tree of life. *American Journal of Botany* 96: 1551–1570.

Young, A. D., and J. Gillung. 2020. Phylogenomics—principles, opportunities and pitfalls of big-data phylogenetics. *Systematic Entomology* 45: 225–247.

Zhang, C., M. Rabiee, E. Sayyari, and S. Mirarab. 2018. ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* 19: 153.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Appendix S1. Voucher information for *Brunellia* species and outgroups included in the phylogenetic analyses.

Appendix S2. Scorings of six morphological characters and their states.

Appendix S3. Gene recovered in each taxon. The value (shading) in each cell indicates gene recovery for that taxon.

Appendix S4. Sequence recovery statistics from Hyb-Seq data for Brunelliaceae and eight taxa included as outgroups. Number of reads (NumReads), mapped to the targets (ReadsMapped), percentage of reads on target (PctOnTarget), number of genes with reads (GenesMapped), number of genes with contigs (GenesWithContigs), wrong percentage of read on target calculated (GenesWithSeqs_, GenesAt25pct, GenesAt50pct, GenesAt75pct, GenesAt150pct), and number of genes with paralogs (ParalogWarnings) are shown here.

Appendix S5. Number of copies retrieved for each locus, by sample, and number of potential paralogs. Blank spaces correspond to one gene to aid visualization.

Appendix S6. Parameters evaluated to minimize potential bias and error in phylogenetic reconstruction using exons. The loci included in the study are shown.

Appendix S7. Parameters evaluated to minimize potential bias and error in phylogenetic reconstruction using supercontigs. The loci included in the study are shown.

Appendix S8. Species tree for exons and supercontig using dataset with and without locus exclusion. Trees were inferred using ASTRAL analysis and maximum likelihood analysis of concatenated data.

Appendix S9. Phylogenetic signal expressed as average bootstrap support of branches within individual gene trees based on (A) exons, and (B) supercontigs. Blue area includes loci with higher than 60% average bootstrap support.

Appendix S10. Concordance factor statistics of the exons dataset with locus exclusion and inferred using maximum likelihood analysis of concatenated data. ID numbers according to branches of the tree below.

Appendix S11. Concordance factor statistics of the supercontig dataset with locus exclusion and inferred using maximum likelihood analysis of concatenated data. ID numbers according to branches of the tree below.

Appendix S12. Species tree for supercontigs with locus exclusion inferred using maximum likelihood analysis of concatenated data. Outgroups were removed to show the relative branch lengths among *Brunellia* species.

Appendix S13. Bar-plots showing concordance of genes (gCF, A) and sites (sCF, B), and discordance gene gDF1, gDF2, gDFP) and sites (sCF, sDF1, sDF2) on exons dataset with locus exclusion, inferred using maximum likelihood analysis of concatenated data. Numbers of branches (ID) according to Appendix S10. The ID numbers were arranged by clade and by their position in the tree (backbone or terminal nodes).

Appendix S14. Significant test for hybridization using concatenated supercontigs matrix with loci excluded. Only those triplets that are significant after Bonferroni correction of P-values (Z-score >4879) are shown. Frequency of predicted hybridization events are shown for each taxon, identified as sp1 to sp57.

Appendix S15. Density plot of gamma values from 4105 hypothesis test for hybridization.

Appendix S16. Infrageneric classification of *Brunellia* according to Cuatrecasas (1985) and Orozco (2001).

How to cite this article: Murillo-A., J., J. Valencia-D., C. I. Orozco, C. Parra-O., and K. M. Neubig. 2022. Incomplete lineage sorting and reticulate evolution mask species relationships in Brunelliaceae, an Andean family with rapid, recent diversification. *American Journal of Botany* 109(7): 1139–1156. https://doi.org/10.1002/ajb2.16025