More-curated Data Outperforms More Data: Treatment of Cryptic and Known Paralogs Improves Phylogenomic Analysis and Resolves a Northern Andean Origin of *Freziera* (Pentaphylacaceae)

Laura Frost[1,2] and Laura Lagomarsino[1,3]

[1]*Department of Biological Sciences and Shirley C. Tucker Herbarium, Louisiana State University, Baton Rouge, LA 70808*

[2] Email: lfrost@lsu.edu

[3] Email: llagomarsino1@lsu.edu

Abstract.—The Andes mountains in South America are a biodiversity hotspot within a hotspot, the New World Tropics, for seed plants. Much of this diversity is concentrated at middle-elevations in cloud forests, yet the evolutionary patterns underlying this extraordinary diversity remain poorly understood. This is partially due to a paucity of resolved phylogenies for cloud forest plant lineages: the young age of the Andes and generally high diversification rates among Andean systems precludes robust phylogenetic inference, and remote populations, few genomic resources, and generally understudied organisms make acquiring high-quality data difficult. We present the first phylogeny of *Freziera* (Pentaphylacaceae), an Andean-centered, cloud forest radiation with potential to provide insight into some of the abiotic and extrinsic factors that promote the highest diversity observed on the globe. Our dataset, representing data for 50 of the ca. 75 spp. obtained almost entirely from herbarium specimens via hybrid-enriched target sequence capture with the universal bait set Angiosperms353, included a proportion of poorly assembled loci likely representing multi-copy genes, but with insufficient data to be flagged by paralog filters: cryptic paralogs. These cryptic paralogs likely result from limitations in data collection that are common in herbariomics combined with a history of genome duplication and are likely common in other plant phylogenomic datasets. Standard empirical metrics for identifying poor-quality genes, which typically focus on filtering for genes with high phylogenetic informativeness, failed to identify problematic loci in our dataset where strong but inaccurate signal was a greater problem. Filtering by bipartition support was the most successful method for selecting genes and resulted in a species tree with lower discordance, higher support, and a more accurate topology relative to a consensus tree. Using known paralogs, we investigate the utility of multi-copy genes in phylogenetic inference and find a role for paralogs in resolving deep nodes and major clades, though at the expense of gene tree concordance and support. With

the first phylogeny, we infer the biogeographic history of *Freziera* and identify the northern Andes as a source region. We also identify distinct modes of diversification in the northern and central Andes, highlighting the importance of fine-scale biogeographic study in Andean cloud forest systems.

Keywords: Angiosperms353; gene tree discordance; gene tree estimation error; environmental filtering; herbariomics; locus filtering; Neotropical biogeography

The Neotropics, the land area between tropical latitudes in the Americas, are potentially home to more seed plant species than the tropical areas of Africa, Asia, and Oceania combined (Humboldt and Bonpland 1807; Humboldt 1808; Gentry 1982; Davis et al. 1997; Myers et al. 2000; Kier et al. 2009; Antonelli and Sanmartín 2011; Raven et al. 2020). Within the Neotropics, the Andes mountains in South America serve as a center of diversity for many lineages and support a significant portion of Neotropical diversity (Gentry 1982; Braun et al. 2002; Mutke and Barthlott 2005; Jørgensen et al. 2011; Mutke and Weigend 2017). As is common in mountain systems globally, Andean species richness exhibits a hump-like distribution, with species richness peaking at mid-elevations (ca. 1500 m;(Rahbek 1995; Kromer et al. 2005; Sang 2009; Guo et al. 2013; Salazar et al. 2015; Quintero and Jetz 2018). These mid-elevation moist forests typically correspond to tropical montane cloud forest, especially in the northern Andes (Hostettler 2002). Gentry (1982) concluded that the explanation for the much greater diversity in the Neotropics lay in understanding diversification patterns in epiphyte, palmetto, and understory shrub lineages of montane forests in the Andes, as these comprised the bulk of taxonomic diversity and seemed to represent rapid radiations. Despite centuries of study, and recent decades of phylogenetic research, Neotropical and Andean diversity remains poorly described and understood (Ulloa et al. 2004; Hopkins 2007; Goodwin et al. 2015; Mutke and Weigend 2017; Zizka et al. 2018; Lagomarsino and Frost 2020). Thus, elucidating evolutionary patterns in Andean-centered cloud forest lineages remains a key step toward understanding the disparity in species richness between the Neotropics and other tropical ecoregions.

The heterogenous and geodiverse landscapes of the Andes, like mountains globally, play a role in generating the biodiversity they house (Ricklefs et al. 1999; Braun et al. 2002; Parks and Mulligan 2010; Antonelli et al. 2018; Hazzi et al. 2018; Flantua et al. 2019; Muellner-Riehl

2019; Muellner-Riehl et al. 2019). The high elevation of Andean mountains forms barriers to wind, creating rainshadows and other localized climatic effects, while their high-relief terrain creates elevational zones and different hydrologic conditions along slopes (Gentry 1982). This results in a mosaic of microhabitats that promote ecological opportunity, small population sizes, and separation between populations leading to speciation (Muellner-Riehl et al. 2019). Through time, orogenic events expand available niches and create new ones for colonization (e.g., the emergence of high alpine grasslands in the Andes--paramo and puna--within the last 5-10 Ma). Climatic fluctuations (e.g., Quaternary glaciation cycles; Flantua et al. 2019) can serve to promote speciation during periods of biome fragmentation and reduce extinction rates during periods of biome connectivity. Mountain systems themselves thus promote parapatric and allopatric speciation. In the Andes, this has been relatively recent: the central Andes rose from nearly half of their current elevation to their present heights within the last 10 Ma (Garzione et al., 2008; Martínez et al., 2020), and the northern Andes are even younger, achieving proportional development in the Pliocene (Gregory-Wodzicki, 2000; Hoorn et al., 2010).

The Neogene uplift of the Andes combined with high diversification rates has resulted in many recent and rapid radiations among Andean lineages. Because of this, phylogenies for these groups are difficult to infer. Short divergence times between speciation events, incomplete lineage sorting, incipient speciation, and introgression have all contributed to poor phylogenetic resolution and high discordance between gene trees and species trees (Vargas et al. 2017; Morales-Briones et al. 2018a). This is further complicated by repeated whole genome duplication events throughout the evolutionary history of plants, at both deep (Jiao et al. 2011; Mayrose et al. 2011; One Thousand Plant Transcriptomes Initiative 2019) and shallow (Chester et al. 2012; Salman-Minkov et al. 2016) scales.There are additional practical limitations for phylogenetic

inference in Andean systems. First, it is difficult to achieve full taxonomic sampling since species are often distributed in remote locations, and members of clades occur in many countries, each with different requirements and restrictions for collection and export permits. As a result, achieving dense taxonomic sampling of Andean-centered lineages commonly requires the use of herbarium specimens as a source of genetic material. Second, collecting conditions--either wet climates or remote locations or both-- delay drying times of specimens, which is detrimental to the preservation of DNA (Brewer et al. 2019). Therefore, tropical herbarium specimens often provide poor-quality DNA (Särkinen et al. 2012; Bakker et al. 2015; Brewer et al. 2019). Improvements in methodology in the past decade bring us closer to achieving resolved phylogenies in these groups. Advancements in genomic sequencing, including hybrid-enriched target sequence capture, allow for the collection of hundreds to thousands of loci (Hart et al. 2016), even from degraded DNA of herbarium specimens (Bakker 2017; McKain et al. 2018). Until recently, development of probesets for target sequence capture required genomic resources in close relatives of the focal system– resources that are often lacking in Neotropical lineages. However, development of universal probe sets for plants (e.g., ferns: (Wolf et al. 2018); flagellate plants: (Breinholt et al. 2021); and angiosperms:(Johnson et al. 2019)) facilitates sequencing of hundreds of loci for any system, regardless of genomic resources available. This provides an opportunity to improve phylogenetic resolution in understudied systems, including Andean plant clades. Additionally, analytical methods are increasingly able to accommodate many biological sources of gene tree discordance, many that are common in phylogenomic datasets of Andean plant clades, including incomplete lineage sorting (ILS:(Ogilvie et al. 2017; Zhang et al. 2018), introgression (Solís-Lemus et al. 2017; Blischak et al. 2018), and gene duplication and loss (GDL:(Molloy and Warnow 2020; Zhang et al. 2020).

Despite these significant advances in the field, challenges to phylogenomics in Andean plant radiations persist, especially those related to paralogy. Prior to methods that could accommodate multi-copy genes, paralogous loci rendered portions of datasets unusable due to assumptions of single-copy orthologs by most phylogenetic methods. Under these assumptions, it is standard for entire loci or their additional copies to be excluded from analyses. While methods have been available to extract orthologs from transcriptomes (Yang and Smith 2014), genomes (Emms and Kelly 2015), and proteomes (Cosentino and Iwasaki 2019; Emms and Kelly 2019), adequate detection of paralogs is required in target capture datasets to filter orthologs (Morales-Briones et al. 2020). This can be difficult if the region has undergone differential loss (Smith and Hahn 2021a, 2021b), resulting in pseudo-orthologs, in which a single copy of the locus is present in each sample despite non-orthology (Koonin 2005). This difficulty can be further exacerbated by artifacts of data collection and assembly that may increase the number of loci with an undetected history of GDL (i.e., cryptic paralogs). For example, in herbariomic datasets, differential success in amplification of degraded DNA, relatively high amounts of missing data, and low coverage and short contigs resulting from generally shorter, lower quality reads relative to fresh tissue could all magnify the problem of cryptic paralogs (Johnson et al. 2016; Gardner et al. 2020). Problems may be further increased if using a universal bait set, as lower specificity of probes for the focal system leads to a higher proportion of off-target amplicons, including additional copies of target genes (Hart et al. 2016; Johnson et al. 2016, 2019; Liu et al. 2019; Gardner et al. 2020). Hidden paralogs are increasingly acknowledged as a source of error in genomic datasets (Smith and Hahn 2021a, 2021b). While coalescent methods that model ILS appear robust to pseudo-orthologs (Markin and Eulenstein 2020; Legried et al. 2021; Smith and Hahn 2021b; Yan et al. 2021), the effect of cryptic paralogs, which do not

represent any biological process, is unknown. As the number of herbariomic target sequence capture datasets continues to rapidly rise and we begin to explore the utility of multi-copy genes for phylogenetic inference (Gardner et al. 2020), we should be careful that the benefit of known paralogs is not lost in the noise of cryptic paralogs.

Researchers are still building evidence for best practices for gene filtering with phylogenomic datasets (Lanier et al. 2014; Liu et al. 2015; Huang and Knowles 2016; Irisarri and Meyer, 2016; Meiklejohn et al. 2016; Simmons et al. 2016; Longo et al. 2017; Molloy and Warnow 2018; Nute et al. 2018). While summary methods (e.g., ASTRAL) appear to be robust to missing data, they may be vulnerable to gene tree estimation error (GTEE; (Molloy and Warnow 2018). Since there is no explicit measurement of GTEE in empirical data, other metrics have been suggested to assess the quality of genes for phylogenetic inference, usually related to phylogenetic signal or informativeness (e.g. alignment length (Liu et al. 2015), number of parsimony informative sites (Leaché et al. 2014), tree length (Smith et al. 2018), the proportion of internal branches in tree length (Shen et al. 2016), and average bootstrap support across all nodes (Blom et al. 2017). However, most of these metrics assume data are accurate and filtering becomes an exercise in separating “weak” genes from “strong” genes (Liu et al. 2015). In herbariomic datasets, the process may be more akin to separating well assembled data from poorly assembled data, such as cryptic paralogs. As a growing number of systems report high discordance among gene trees (Degnan and Rosenberg 2009; Wickett et al. 2014; Copetti et al. 2017; Vargas et al. 2017, 2019; Morales-Briones et al. 2018a, 2018b, 2021; Pease et al. 2018; Liu et al. 2019), it is worth examining how different criteria for gene selection in these large datasets impact topology, discordance and support, especially with herbariomic datasets that may include a higher proportion of low quality data. While the addition of loci to phylogenetic

analyses can improve resolution (Streicher et al. 2016), trade-offs have been observed between the amount of data and the quality of data (e.g., locus length, number of informative sites, and model fit; Degnan and Rosenberg 2009; Reid et al. 2014; Liu et al. 2015; Xu and Yang 2016). Empiricists may be hesitant to exclude data that was time- and cost-intensive to collect, but the benefit of more data relative to more-curated data should continue to be examined in empirical systems.

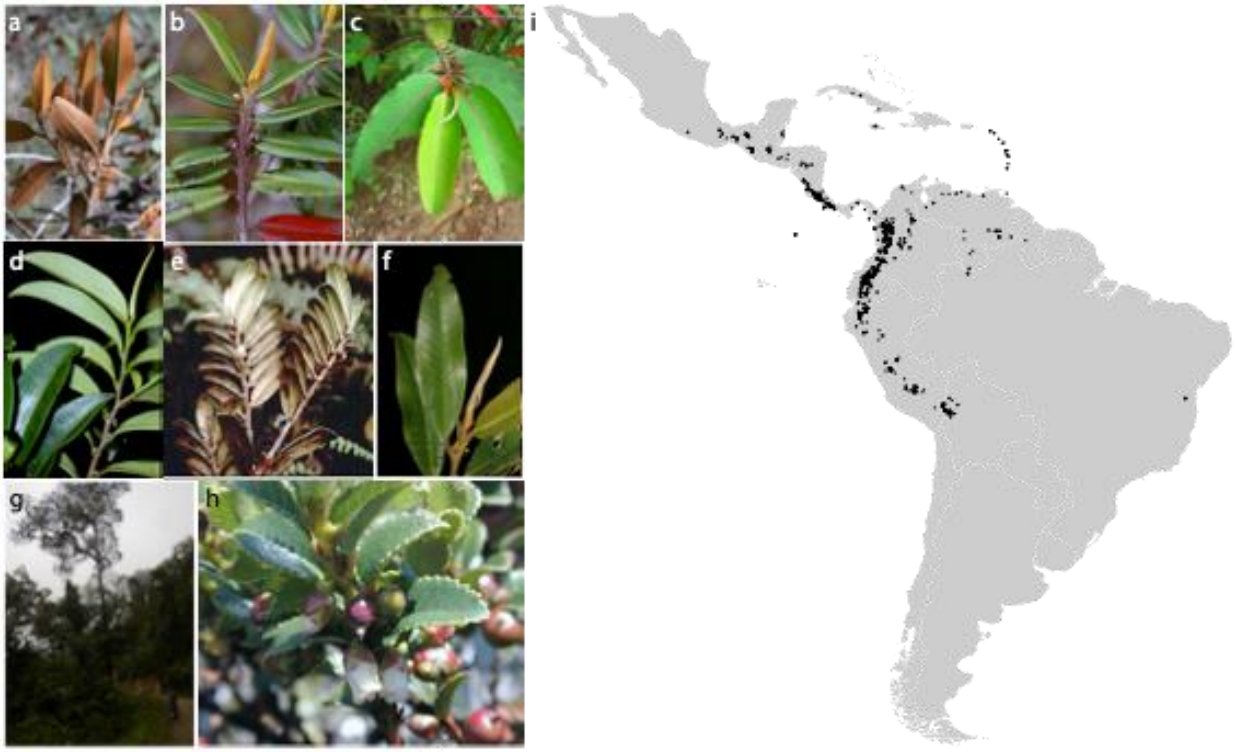
Difficulty generating well-resolved phylogenies is not the only limiting factor for understanding patterns of Andean cloud forest diversification; the multitude of biotic and abiotic factors contributing to diversification make it difficult to attribute patterns to their generating process. One of the notable drivers of diversification in Andean-centered lineages is specialized pollination systems (Gentry 1982; Lagomarsino et al., 2016). High diversity of pollinators, both in species number and guild (e.g., bat, bee, bird, moth), frequent shifts between guilds of pollinators, and tight mutualisms within guilds have allowed differentiation between populations within the same geographic area. Other key innovations (e.g., epiphytism) also drive diversification in cloud forests (Gentry and Dodson 1987; Gravendeel et al. 2004; Givnish et al. 2014, 2015; Donoghue and Sanderson 2015; Muellner-Riehl et al. 2019). Though these intrinsic factors may represent an evolutionary theme among Andean cloud forest lineages, they confound the extrinsic factors of the Andes that further promote diversification. Lineages boasting floral diversity and high diversification rates undoubtedly provide valuable insight into diversification dynamics in Andean cloud forests and are attractive to evolutionary biologists hoping to understand extraordinary biodiversity (Beaulieu and O'Meara 2018). However, the extrinsic factors that have also contributed significantly to producing the highest biodiversity in the world remain understudied in cloud forest lineages. Understanding how the Andes themselves

influence evolutionary patterns in cloud forest plants will help to understand what sets the Andes and the Neotropics apart from other global mountain chains and ecoregions, respectively.

Freziera, a genus of 75 spp. of Neotropical trees and shrubs, can help illuminate patterns associated with extrinsic factors in evolution of cloud forest biota. *Freziera* is widely distributed throughout the mountane regions of the Neotropics, from southern Mexico to Bolivia, with a center of diversity in the Andes (61 spp. are distributed in Andean cloud forests;(Santamaría-Aguilar and Monro 2019); Fig. 1). Species are mostly restricted to cloud forests (≥1000 m a.s.l.), but range in elevation within the mid-elevation, moist forest biome (Weitzman 1987, 1988; Santamaría-Aguilar and Monro 2019). Unlike many charismatic cloud forest radiations, *Freziera* species are consistent in life history traits related to reproduction: species are dioecious (i.e., separate staminate (pollen producing) and pistillate (ovule producing) individuals), flowers exhibit a generalist invertebrate pollination syndrome (i.e., small, pale green to white flowers with a narrow opening for an insect proboscis), and all species produce fleshy berries as fruits (Weitzman 1987, 1988; Weitzman et al. 2004; Santamaría-Aguilar and Monro 2019). They do not have large, showy flowers or particularly variable morphology in flowers or fruit. Instead, concomitant with this range of elevational distribution is a high degree of variation in leaf traits (Fig. 1). Shape ranges from orbicular to elongate; indument varies in presence, color, and texture; and, most notably, leaf size varies over 400-fold within the genus. Leaf morphology has well-established correlations with the environment due to leaves' role in photosynthesis, respiration, and transport of water, nutrients, and photosynthates (Ivey and DeSilva 2001; Givnish 2008; Oguchi et al. 2018). The diversity of leaf shapes and sizes observed in *Freziera* suggests a degree of environmental adaptation among species. Identifying patterns of niche differentiation in *Freziera* will shed light on the role of ecological opportunity in Andean

diversification. Meanwhile, the widespread distribution in the Andes and other Neotropical mountain ranges (e.g., the Talamanca range in Costa Rica) suggests dispersal has also been a factor in diversification. The evidence of dispersal and adaptation makes *Freziera* a good system to understand how an Andean cloud forest lineage lacking some of the more obvious biotic drivers of diversification— an Andean "minivan", to borrow a phrase from Beaulieu and O'Meara (2018) — moves and evolves.

FIGURE 1.



Though there are many challenges toward phylogenetic inference in Andean clades, these phylogenies are a fundamental step toward understanding the evolutionary patterns that contribute to the distribution of biodiversity on the globe. We generate the first phylogeny of *Freziera* using data almost entirely collected from herbarium specimens with the universal probe set Angiosperms353 (Johnson et al., 2019). As an herbariomic dataset that includes identifiable paralogs for an Andean radiation, *Frezeria* is an appropriate system to understand how locus

filtering and the inclusion of paralogous loci impact gene tree-species tree discordance, species tree resolution, and support. With a time-calibrated phylogeny, we infer the biogeographic history of *Freziera* and quantify how species are distributed in environmental niche space. Despite limitations in data collection that come with understudied Neotropical systems, we infer a robust phylogenetic hypothesis for the Neotropical radiation *Freziera* and uncover distinct modes of diversification across regions within the montane cloud forest biome.

MATERIALS AND METHODS

Taxon Sampling

Ninety-three accessions representing 55 *Freziera* species—approximately 73% of the taxonomic diversity of the genus—were sampled for the in-group. Nine accessions from other genera in Pentaphylacaceae were sampled for the outgroup, including *Eurya japonica*, from the genus sister to *Freziera*; *Cleyera albopunctata,* another member of the same tribe, Frezierieae; and 7 species of *Ternstroemia,* belonging to the tribe sister to Frezierieae: *T. candolleana, T. peduncularis, T. sp., T. stahlii, T. gymnanthera, T. pringlei,* and *T. tepezapote* (Weitzman et al. 2004; Tsou et al. 2016).

DNA Extraction, Library Preparation, Target Enrichment, and Sequencing

Five hundred mg of dried leaf tissue, primarily from herbarium specimens, was homogenized using a FastPrep-24TM 5G bead beating and lysis system (MP Biomedicals, Solon, Ohio, United States). DNA extraction followed a modified sorbitol extraction protocol (Štorchová et al. 2000). Double-stranded DNA concentration was quantified using a Qubit 4 Fluorometer (Invitrogen, Waltham, Massachusetts, United States) and fragment size was

assessed on a 1% agarose gel. For samples with a high concentration of large fragments (>800 bp), DNA was sheared using a Bioruptor Pico (Diagenode Inc., Denville, New Jersey, United States) until most fragments were less than 500 bp in length.

Library preparation was carried out using KAPA Hyper Prep and KAPA HiFi HS Library Amplification kits (F. Hoffmann-La Roche AG, Basel, Switzerland) and with iTru i5 and i7 dual-indexing primers (BadDNA, University of Georgia, Athens, Georgia, United States). Library preparation with KAPA Hyper Prep followed the manufacturer's protocol (KR0961 – v8.20) with the following modifications: reaction volumes were halved (i.e., 25 μL starting reaction) and bead-based clean-ups were performed at 3X volume rather than 1X volume to preserve more small fragments from degraded samples that are characteristic of herbarium specimens. As the 3X volume bead-based clean-up retains adapter dimers as well as short fragments, samples were visualized again using a 1% agarose gel to identify samples with many fragments less than 100 bp long. Those samples were processed with a GeneRead Size Selection kit to remove fragments shorter than 150 bp (Qiagen, Germantown, Maryland, United States). Library amplification reactions were performed at 50 μL.

Target enrichment was carried out using the MyBaits Angiosperms353 universal probe set (Däicel Arbor Biosciences, Ann Arbor, MI;(Johnson et al. 2019). Target enrichment followed the modifications to the manufacturer's protocol outlined in Hale et al. (2020; i.e., pools of 20-24 samples and RNA baits diluted to ¼ concentration). Twenty nanograms of unenriched DNA library were added to the cleaned, target enriched pool to increase the amount of off-target, chloroplast fragments in the sequencing library. DNA libraries were sent to Novogene Corporation Inc., (Sacramento, California, United States) for sequencing on an Illumina Hiseq 3000 platform (San Diego, California, United States) with 150 bp paired-end reads.

Raw Data Processing and Locus Extraction and Alignment Cleaning

Raw sequence reads were demultiplexed by Novogene Corporation Inc., (Sacramento, California, United States). Adapter sequence removal and read trimming were performed using illumiprocessor v2.0.9 (Faircloth 2013, 2016), a wrapper for trimmomatic v0.39 (Bolger et al. 2014). The default settings were used and reads with a minimum length of 40 bp kept.

Trimmed reads were assigned to their target genes, assembled, and aligned using HybPiper v1.3.1 (Johnson et al. 2016). Read mapping and contig assembly were performed using the reads_first.py script. The intronerate.py script was run to extract introns and intergenic sequences flanking targeted exons. Coding and non-coding regions were extracted using the retrieve_sequences.py script with “dna” and “supercontig” arguments, respectively. Supercontigs include both coding and non-coding regions as a single concatenated sequence for each target gene. Individual genes were aligned using MAFFT v. 7.310 (Katoh and Standley 2013). Thirty-one loci were flagged as paralogous by the paralog_retriever.py script, and either removed downstream analyses or treated explicitly as paralogs.

Phylogenetic Analyses

Gene tree inference and filtering.— Preliminary gene trees were generated from aligned sequences for the 322 loci lacking paralog flags with RAxML v8.2.12 (Stamatakis 2014) under the GTR model with optimization of substitution rates and site-specific evolutionary rates (-m GTRCAT) and 200 rapid bootstrap replicates. The preliminary trees were then processed with TreeShrink v1.3.3 (Mai and Mirarab 2018) on a “per-gene” and “all-gene” basis to identify long branches that are likely associated with spurious sequences. The “per-gene” test identifies

exceedingly long branches from the distribution of signature values (i.e., the maximum reduction in tree diameter resulting from removal of a set of terminal branches) within each gene, whereas the “all-gene” creates one distribution based on all genes to which species signatures are compared (Mai and Mirarab 2018). The identified samples were removed from alignments, except in instances where the entire *Ternstroemia* clade in the outgroup was identified, as this was considered more likely to reflect clade-specific differences in mutation rate and/or time elapsed since MRCA than spurious alignments. Summary statistics for alignments were obtained using AMAS (Borowiec 2016), including alignment length, number of variable sites, and proportion of parsimony informative sites. Genes with fewer than 25 ingroup samples were excluded from further analyses. Gene trees were inferred with IQ-TREE multicore v2.1.1 (Nguyen et al. 2015; Minh et al. 2020) combining model section via ModelFinder (Kalyaanamoorthy et al. 2017), tree search, 5000 ultrafast bootstrap replicates (Hoang et al. 2018), and 5000 Shimodaira-Hasegawa-like approximate likelihood-ratio tests (SH-aLRT;(Guindon et al. 2010; Anisimova et al. 2011)). Average bootstrap support and percent of tree length made up by internal branches were extracted from the IQtree output for each gene tree.

Genes and gene trees were filtered using ten different empirical criteria to exclude potential sources of gene tree estimation error. Alignment length (>1000 bp; Liu et al. 2015) was used on the assumption that longer genes are more phylogenetically informative than shorter genes. Loci were also filtered by the number of variable sites, proportion of parsimony informative sites, and the number of parsimony informative sites relative to the number of unrooted internal branches (Number of parsimony informative sites/(number of tips-3)). For alignment-based metrics, outgroup sequences were excluded and summary statistics were

calculated for the ingroup using AMAS. Thresholds near average were selected for number of variables sites (>750), proportion of parsimony informative sites (0.075), and number of parsimony informative sites relative to the number of branches (4x number of internal branches) to select high-quality loci by those metrics while maintaining a similar number of loci for each criterion (i.e.,~150 loci).

Tree length can reflect the amount of molecular evolution in a gene--longer trees indicate more substitutions--and, therefore, may reflect phylogenetic informativeness. Gene trees were filtered for those with above average tree length. Additionally, because internal branch lengths have been shown to be correlated with phylogenetic signal (Shen et al., 2016), gene trees were filtered based on the percent of tree length made up by internal branches. While a high percentage of internal branch lengths can signal good resolution between samples, long internal branch lengths can also signal pseudo-orthologs (Smith and Hahn, 2016b)--and, likely, cryptic paralogs. Therefore, both sets of gene trees with above average percentages of internal branch lengths and below average percentages of internal branch lengths were examined. As measures of gene tree uncertainty, gene trees were filtered for those with above average bootstrap support across all nodes. Lastly, gene trees were filtered based on above average bipartition support compared to a species tree and below average root-to-tip variance. These last two metrics were calculated with SortaDate (Smith et al. 2018), including phyx (Brown et al. 2017). Gene trees were rooted using pxrr in phyx, and the bipartition support and root-to-tip variance were calculated using scripts get_bp_genetrees.py and get_var_length.py, respectively. The latter script additionally calculates tree length, which can be interpreted as the amount of variation/molecular evolution in the gene. The SortaDate calculations were combined using the combine_results.py and “good” genes ranked via get_good_genes.py scripts. Bipartition support

was calculated against an ASTRAL-III tree (Zhang et al. 2018) inferred from the best maximum likelihood (ML) topologies of 314 gene trees. With the assumption that coalescent methods are robust enough to noise to produce an approximate species tree from all supposed orthologous loci (Markin and Eulenstein 2020; Legried et al. 2021; Smith and Hahn 2021b; Yan et al. 2021), we expect that truly single-copy gene trees should be more similar to the species tree than trees for cryptic multi-copy genes. For the first level of filtering, gene trees with average and above bipartition support were selected. The second level of filtering selected from genes with above average bipartition support genes that additionally had lower than average root-to-tip variance, or greater clock-likeness. This filtering scheme was based on the recommended ranking of metrics in Smith et al. (2018).

Species tree inference.— To improve species tree inference by decreasing the amount of missing data, accessions present in <10% of gene trees were trimmed from gene trees using R package ape (Paradis and Schliep 2019). Maximum likelihood trees generated by IQtree were used to estimate species trees in ASTRAL-III (Zhang et al. 2018) when analyzing only single-copy loci or ASTRAL-Pro (Zhang et al. 2020) when analyzing a combination of single-copy and multi-copy loci. To examine the impact of more data versus more-curated data and the addition of paralogs (i.e., multi-copy genes successfully flagged by HybPiper) on phylogenetic inference, twenty-two datasets were analyzed: one including all 314 loci not flagged as paralogous by HybPiper, each of the ten filtered datasets, and those eleven datasets with the addition of the 31 known paralogous loci. A list of datasets and names applied to them for the remainder of the paper can be found in Table 1.

TABLE 1. Dataset Names, Description, and Number of Loci Included.

Dataset name	Description	Number of loci (+ paralogs)
clocklike.bipartition(+para)	gene trees with above average bipartition support and below average root-to-tip variance	131 (162)
tree.length(+para)	total tree length above average	140 (171)
PI.per.branch(+para)	number of parsimony informative sites at least 4x the number of internal branches	148 (179)
high.%.internal(+para)	proportion of total tree length comprising internal branch lengths above average	149 (180)
proportion.PI(+para)	alignment contains at least 7.5% parsimony informative sites	160 (191)
average.BS(+para)	average bootstrap support for gene tree above average	163 (193)
low.%.internal(+para)	proportion of total tree length comprising internal branch lengths below average	165 (196)
bipartition(+para)	gene trees with above average bipartition support relative to a preliminary species tree	166 (197)
variable.sites(+para)	alignment contains at least 750 variable sites in	168 (199)
1000bp(+para)	alignment at least 1000 bp	300 (331)
all.orthologs(+para)	all loci not flagged as paralogs	314 (345)

Gene tree discordance was assessed for all species tree topologies using the final normalized quartet score (NQS) provided in the ASTRAL output. Discordance was additionally assessed for *all.orthologs*, *bipartition*, and *clocklike.bipartition* using Phyparts (Smith et al. 2015) and visualized with PhypartsPieCharts (https://github.com/mossmatters/phyloscripts/tree/master/phypartspiecharts). Average node support, proportion of well-supported nodes (ppl≥0.95), and RF distances were calculated as summary statistics for species tree inference. A majority-rule consensus tree was generated from the twenty-two resulting species trees using the consensus() function in ape. RF distances were calculated between species trees and the consensus tree using the RFdist() function in R package phangorn v2.6.3 (Schliep, 2011).

Divergence time estimation.—A time-calibrated species tree was estimated with *BEAST v.1.8.4 (Heled and Drummond 2010). Due to the computational limits of the MCMC algorithm in *BEAST, we reduced our full dataset to a subset of loci that had higher than average bipartition support, lower than average root-to-tip variance, and high taxonomic sampling; this

resulted in 19 loci. Starting parameters and priors were set using BEAUTi v.1.8.4 (Drummond et al. 2012). Site model, clock model and partition tree were unlinked for all loci. The analysis was informed by results of the ASTRAL-III analyses of the full data set: major clades identified in the consensus tree were constrained as monophyletic with “Species Sets”, as was the branching order of those clades. A GTR substitution model with a gamma distribution plus invariant site heterogeneity model was applied to each locus, as was an uncorrelated relaxed clock (Drummond et al. 2006). A birth-death prior was set on the species tree analysis with a piecewise linear and constant root population size model. Two secondary calibrations based on a recent fossil-calibrated phylogeny of Ericales (Rose et al. 2018) were applied for time-calibration: 66.7 Ma at the ancestral node of *Freziera* and *Ternstroemia* and 34.6 Ma at the ancestral node of *Freziera* and *Cleyera*. Each tmrca prior was assigned a normal distribution with a standard deviation of 2.5. The uncorrelated lognormal relaxed clock mean was changed from a fixed value of one to a lognormal distribution with an initial value of one; remaining priors were left with their default settings. Four independent runs of 100 million generations, sampling every 25,000 generations, were performed. MCMC trace files were assessed in Tracer v.1.6.0 (Rambaut et al. 2018) to ensure that the runs had converged, reached stationarity, and that effective sample sizes for metrics were greater than 200. Species tree samples from the four runs were combined with LogCombiner v.1.8.4 (Drummond and Rambaut 2007); the maximum clade credibility tree was selected and support for the topology applied with TreeAnnotator v.1.8.4 (Drummond and Rambaut 2007). The number of lineages through time from the dated phylogeny were plotted using the ltt.plot() function in ape.

Biogeographic Reconstruction

Ancestral areas were estimated using BioGeoBEARS (Matzke 2013a, 2013b, 2014; Massana et al. 2015; R Core Team 2017) implemented in RASP v4.0 beta (Yu et al. 2015). The *BEAST topology was trimmed to exclude the outgroup and served as the input tree. Four discrete areas were defined in the Neotropics: (1) Mesoamerica, (2) the Guiana Shield, comprising eastern Venezuela, Guyana, Suriname, French Guiana, and northern Brazil, (3) the Northern Andes, comprising northwestern Venezuela, Colombia, Ecuador, and northern Peru, and (4) the Central Andes, comprising central to southern Peru and Bolivia. *Freziera grisebachii* is additionally distributed in the Caribbean, however, this area was excluded from analyses since it was only part of one species' distribution. The maximum number of areas was set to three. Six biogeographic models--DIVAlike, BayAREAlike, and DEC, as well as the addition of jump dispersal for each of those models--were compared, and the DEC+J model (Matzke 2014) was selected based on AICc scores. The apparent dispersibility of the genus, suggested by its occurrence on islands and deeper Asian ancestry (Rose et al. 2018), supports DEC+J as a plausible model of dispersal. To further visualize the geospatial relationships within subclades, occurrence points pulled from GBIF and the literature were linked to their respective tips on the species tree and plotted on a map of the Neotropics with the phylo.to.map function in R package phytools.

Climate and Soil Data

Latitude and longitude for *Freziera* collections were extracted from georeferenced, databased specimens available on the Global Biodiversity Information Facility (gbif.org). Duplicate localities and country centroids were removed from the extracted data. Because *Freziera* has received a recent taxonomic revision (Santamaría-Aguilar and Monro 2019), data

were not available for all species through GBIF; for these recently described species, locality data were gathered from Santamaría-Aguilar and Monro (2019; *F. golondrinensis*, *F. monteagudoi*, *F. peruana*, and *F. siraensis*) and Cuello and Santamaría-Aguilar (2015; *F. guaramacalana*). Where applicable, points were also removed from the species to which these specimens were previously assigned. Using R package raster (Hijmans et al. 2015), the standard WorldClim 2.0 30s Bioclimatic variable layers (Fick and Hijmans 2017) and the WorldClim 2.1 30s elevation layer (https://www.worldclim.org/data/worldclim21.html) were stacked and clipped to the tropical latitudes of the Americas (extent = -120, -30, -23, 23). Climate data and elevation were extracted and averaged for each species.

Phylogenetic PCA

To summarize environmental variables while accounting for evolutionary relationships, phylogenetic principal component analyses (pPCA) were performed on the species average datasets for climate and soil using the phyl.pca() function in the R package phytools (Revell 2009, 2012). Settings for method and mode were Brownian Motion model (method="BM") and correlation (mode = "corr"), respectively. Principal component (PC) scores from the first two PCs were used as variables in downstream phylogenetic comparative methods.

RESULTS

Taxon Sampling and Target Sequence Capture Success

Two of the 93 *Freziera* accessions (i.e., Freziera_cordata_NZ1539 and Freziera_calophylla_AG16907) failed to amplify for any of the targeted loci. On average, sequences were recovered for 241 loci (348 maximum), and 57 samples from the 102 total

submitted for sequencing recovered more than 300 loci of the 353 targeted. Voucher information for accessions and per sample data for read trimming and contig assembly are available in Supplementary Table 1. Data were recovered for at least one sample in all 353 loci, and at least 25% of samples were present in 346 loci; per locus summary statistics are available in Supplementary Table 2. Seventy-six and 72 accessions representing 50 of the 75 species of *Freziera* were included in the final ASTRAL-III and *BEAST analyses, respectively, after the rounds of processing described (see methods; Supplementary Table 1).

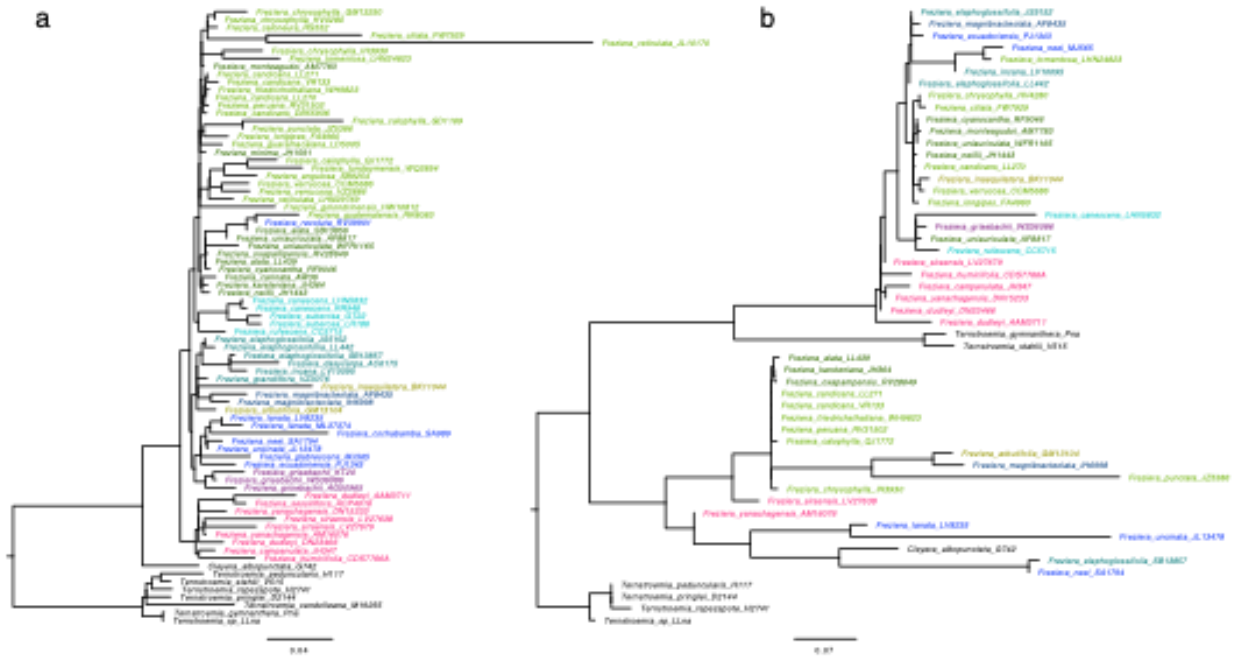
Phylogenetic Analyses

Gene tree inference and filtering.—Of the 322 loci that were not flagged as paralogs by HybPiper, eight were removed because they contained fewer than 25 ingroup samples, leaving 314. Summary statistics for alignments, both including and excluding outgroups, are available in Supplementary Table 2; gene trees are available on Dryad (https://doi.org/10.5061/dryad.jsxksn09g). Dataset names, thresholds applied, and number of loci meeting thresholds are summarized in Table 1. All loci were at least 600 bp (range: 619-12,886 bp)--well above the 100 bp threshold of “weak” genes (Liu et al., 2015); 300 were longer than 1,000 bp, the threshold for “strong” genes. The average number of variable sites for alignments of the ingroup was 884 (range: 111-3,879 sites). The threshold was set at 750 variable sites to increase the number of loci from 131 (i.e., those above average) to 168. Similarly, the average proportion of parsimony informative sites per locus was 0.086, with 125 loci with above average values, but the threshold was set at 0.075 to increase the number of loci included in that dataset to 160. For the number of parsimony informative sites relative to tree size, the number of loci

increased from 118 to 148 when the threshold was relaxed from the average of 4.8 (range: 0.68-26.10) to 4.0 sites per internal branch.

One hundred forty gene trees exhibited tree lengths higher than average (2.48; range: 0.47-8.82), and, on average, total tree length comprised 36.5% (range=17.46-61.53%) internal branch lengths. Gene trees were divided into 149 with above average percent internal branch lengths and 165 with below average internal branch lengths. Average bootstrap support was 63.62, and 163 gene trees with above average support were selected. SortaDate identified 166 gene trees with near average or higher bipartition support (average ICA = 0.0617; threshold ≥0.0612; Supplementary Table 2). Additional filtering by root-to-tip variance resulted in 131 gene trees with high bipartition support and clock-likeness (average and threshold root-to-tip variance=0.021; Supplementary Table 2). Examination of gene trees with lower than average bipartition support found topologies consistent with cryptic paralogy (i.e., gene trees with non-monophyletic species and deep divergences between clades containing separate individuals of non-monophyletic species; Figure 2).

FIGURE 2.



Species tree inference.—Values for normalized quartet score, average node support, proportion of well-supported nodes, and RF distances between species tree topologies and the consensus trees are listed in Table 2. Of the different criteria used to filter orthologs and their gene trees, filtering by bipartition support resulted in the best overall tree (*biparition*; Table 2). Further filtering for low root-to-tip variance (*clocklike.bipartition*) resulted in the highest normalized quartet score (NQS: 0.621; i.e., the lowest discordance among gene trees) and average node support (avg. ppl: 0.800). Filtering by bipartition support alone resulted in similar levels of concordance (NQS: 0.601) and support (avg. ppl: 0.792), as well as the highest proportion of well-supported nodes (0.455 of nodes with ppl≥0.95) and the lowest RF distance between species and consensus trees (RF distance: 14). Aside from being the best performing datasets for species tree inference, these were the only two datasets to outperform *all.orthologs*. Nearly all of the other filtering schemes resulted in poorer quality species trees across all measures of concordance and support. Exceptions include higher NQS for *low.%.internal* (NQS:

0.586 versus 0.566 for *all.orthologs*) and a higher proportion of well-supported nodes for *1000bp* (0.429 of nodes with ppl≥0.95 versus 0.416 for *all.orthologs*).

In all cases the addition of known paralogs increased discordance, and in most cases reduced average node support and/or the proportion of well-supported nodes. On average, including paralogs decreased NQS by 0.01; the greatest reduction in NQS (i.e., the largest increase in discordance) was for the datasets with the highest concordance when paralogous loci were excluded (e.g., *clocklike.bipartition*, *bipartition*, and *low.%.internal*). The addition of paralogs improved both measures of support for *variable.sites* (from avg. ppl=0.746 to 0.754 and from 35.1% of nodes with ppl≥0.95 to 36.4%). Paralogs slightly improved average node support for *tree.length* (from an avg. ppl of 0.697 to 0.703), *PI.per.branch* (0.709 to 0.719), and *1000 bp* (0.773 to 0.775). Lastly, paralogs increased the proportion of well-supported nodes for *low.%.internal* (from 33.8 % of nodes with ppl≥0.95 to 39.0%), *high.%.internal* (24.6% to 27.3%), and *proportion.PI* (27.3% to 28.6%). Despite decreases in concordance and support, the addition of paralogs generally improved topology; eight out of ten datasets showed a decrease in RF distance between species and consensus trees when paralogs were included (Table 2). The greatest reduction in RF distance with the inclusion of paralogs was for *clocklike.bipartition* (from 34 to 20). Tree distance increased slightly with the addition of paralogs for *bipartition* (from 14 to 18), which, along with *1000bp*, had the most similar topology to the consensus tree of datasets including only orthologs. The *all.orthologs+para.* and *1000bp+para.* datasets had the most similar topologies to the consensus tree (RF distance=12) of the datasets including paralogs.

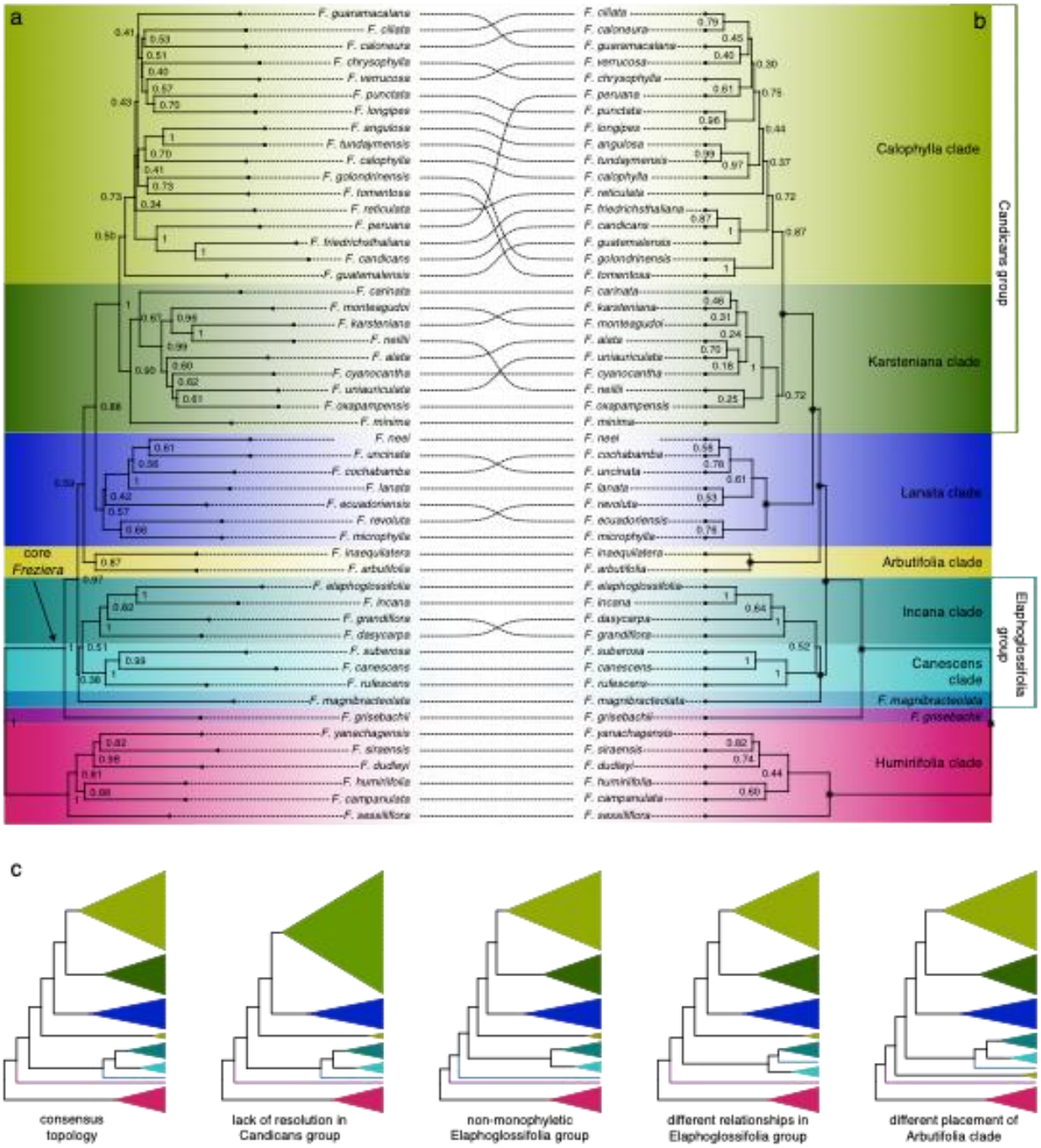
The consensus topology included nine clades that were frequently inferred across analyses: the Humiriifolia clade, *F. grisebachii*, *F. magnibracteolata*, the Canescens clade, the

Incana clade, the Arbutifolia clade, the Lanata clade, the Karsteniana clade, and the Calophylla clade (Fig. 3). Eight datasets produced species trees consistent with the nine clades and the branching order of those clades in the consensus tree: *clocklike.bipartition* (except for *F. minima*), *clocklike.bipartition+para*, *bipartition*, *bipartition+para*, *variable.sites*, *1000bp*, *1000bp+para*, *all.orthologs+para* (Table 2; Supplementary Fig. 1a-b,i-j,m,t,v-w). Among species trees consistent with the consensus topology, some deep nodes--the common ancestor of all *Freziera*, the common ancestor of core *Freziera*, and the successive node within core *Freziera*--and named clades--the Humiriifolia, Canescens, and Incana clades, as well as the Candicans group (Fig. 3)--were inferred with high support (ppl≥0.95) across all analyses.

TABLE 2. Summary of Dataset Performance in Species Tree Analyses

Dataset name	Normalized Quartet Score	Average node support (ppl)	proportion of nodes with ppl≥0.95	RF distance from consensus	Two clades in Candicans group	Monophyletic Elaphoglossifolia group	Branching order of Elaphoglossifolia group	Placement of Arbutifolia clade
clocklike.bipartition(+para)	**0.621 (0.597)**	**0.800 (0.781)**	0.403 (0.364)	34 (20)	-* **(+)**	+ **(+)**	+ **(+)**	+ **(+)**
tree.length(+para)	0.524 (0.521)	0.697 (0.703)	0.273 (0.273)	54 (52)	- (-)	- **(+)**	n/a (-)	- (-)
Pl.per.branch(+para)	0.537 (0.531)	0.709 (0.719)	0.273 (0.273)	54 (46)	- (-)	- **(+)**	n/a (-)	+ **(+)**
high.%.internal(+para)	0.541 (0.533)	0.693 (0.690)	0.247 (0.273)	56 (50)	- (-)	+ **(+)**	+ (-)	+ **(+)**
proportion.Pl(+para)	0.527 (0.523)	0.718 (0.714)	0.273 (0.286)	44 (50)	- (-)	- **(+)**	n/a (-)	+ **(+)**
average.BS(+para)	0.560 (0.550)	0.727 (0.723)	0.312 (0.286)	24 (20)	+ **(+)**	+ **(+)**	- (-)	+ **(+)**
low.%.internal(+para)	**0.586 (0.571)**	0.760 (0.748)	0.338 **(0.390)**	34 (26)	+ (-*)	- **(+)**	n/a **(+)**	- (-)
bipartition(+para)	**0.601 (0.584)**	**0.792 (0.785)**	**0.455 (0.416)**	**14 (18)**	+ **(+)**	+ **(+)**	+ **(+)**	+ **(+)**
variable.sites(+para)	0.559 (0.550)	0.746 (0.754)	0.351 (0.364)	28 (24)	+ **(+)**	+ (-)	+ (n/a)	+ (-)
1000bp(+para)	0.566 (0.559)	0.773 **(0.775)**	**0.429 (0.390)**	**14 (12)**	+ **(+)**	+ **(+)**	+ (+)	+ **(+)**
all.orthologs(+para)	0.566 (0.559)	**0.777 (0.775)**	**0.416 (0.390)**	**16 (12)**	+ **(+)**	- **(+)**	n/a **(+)**	+ **(+)**

FIGURE 3.



Support varied along the backbone of core *Freziera* and for some clades within. Regions of low support also tended to be sources of conflict between other species trees and the consensus topology. *Freziera magnibracteolata*, the Canescens clade, and the Incana clade

formed a larger clade, named the Elaphoglossifolia group (Fig. 3). For the eight species trees consistent with the consensus topology, the Canescens and Incana clades were inferred with full support (ppl=1.0, except ppl=0.99 for the Incana clade in the *variable.sites* tree). However, relationships between subclades of the Elaphoglossifolia group lacked support ((Canescens clade, Incana clade) ppl: 0.49-0.65, *F. magnibraceolata*) ppl: 0.31-0.39; Supplementary Fig. 1). Six other datasets also inferred a monophyletic Elaphoglossifolia group but with a different branching order from the consensus topology and similarly low support ((Incana clade, *F. magnibracteolata*) ppl: 0.42-0.70, Canescens clade) ppl: 0.35-0.47). The remaining six datasets resulted in a non-monophyletic Elaphoglossifolia group (Table 2), but each of its three subclades were still recovered. The Candicans group, comprising the Calophylla and the Karsteniana clades, on the other hand, was always inferred (and generally well-supported; ppl≥0.95 in 18 of the 22 species trees), whereas its subclades were not. Neither subclade was strongly supported in any analysis; however, the Karsteniana clade tended to have weak-to-moderate support (ppl: 0.40-0.93), while the Calophylla clade had weak support (ppl: 0.35-0.75). In most cases in which the two clades were not resconstructed (Table 2), members of the Karsteniana clade still formed a monophyletic group and members of the Calophylla clade were paraphyletic with respect to the Karsteniana clade. However, in the species tree for *proportion.PI*, the Karsteniana clade was polyphyletic within the Candicans group, and *clocklike.bipartition* did find the two clades except that *F. minima*, usually reconstructed as sister to the rest of the Karsteniana clade, nested within the Calophylla clade.

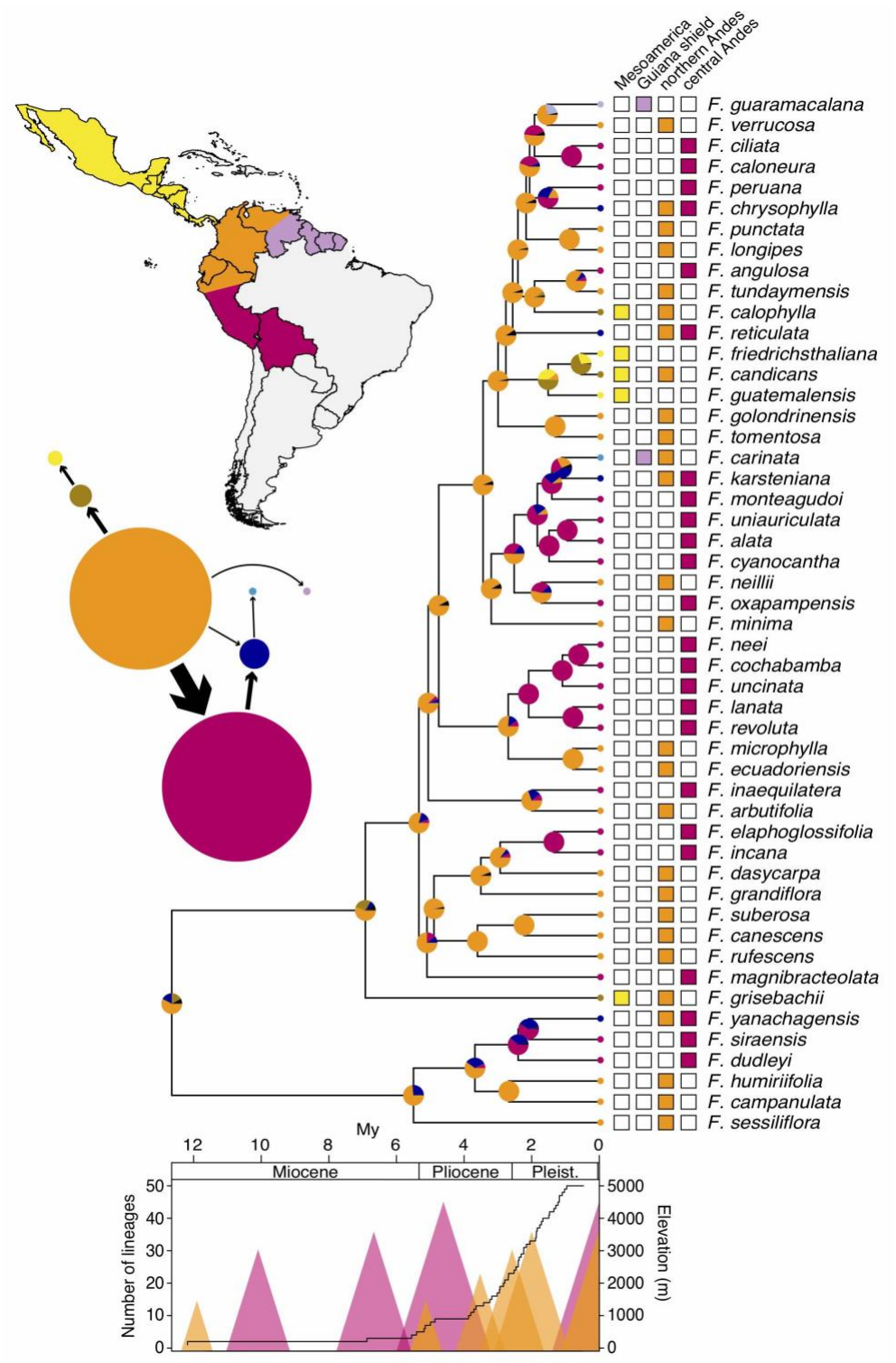
Other areas of weak support include the Lanata and Arbutifolia clades as well as the placement of the Arbutifolia clade. The Lanata clade was consistently inferred across all analyses, albeit with poor support (ppl: 0.32-0.87), as was the placement of the Lanata clade as

sister to the Candicans group with generally moderate support (ppl: 0.65-1.0, but ppl≥0.95 for 10 out of 22 analyses). Similarly, the Arbutifolia clade was weakly supported (ppl: 0.58-0.94) but was recovered in every analysis. The relationship of the Arbutifolia clade as sister to the Lanata clade and the Candicans group was inferred by most analyses, if weakly supported (ppl: 0.48-0.85). However, *low.%.internal*, *low.%.internal+para.*, and *variable.sites+para.* datasets recovered it in a grade with *F. grisebachii* and sister to the rest of core *Freziera*; it came out as sister to *F. grisebachii* in species trees for *tree.length* and *tree.length+para*.

Assessment of gene tree concordance with PhyParts revealed high levels of conflict between gene trees and the species tree topology based on the dataset including all supposed orthologs and paralogs. Most of the conflict was the result of many different alternative topologies rather than one frequent alternative topology (Supplementary Fig. 2).

Divergence time estimation.—Relationships at unconstrained nodes in the time-calibrated species tree estimated in *BEAST were consistent with the *all.orthologs+para* for the Humiriifolia and Canescens clades (Fig. 3a,b), which were recovered with full support by every dataset. Clades with short divergence times or few coalescent units between speciation events in the *BEAST and the ASTRAL-III tree, respectively (e.g., the Candicans group and the Lanata clade), tended to have the most disagreement in species relationships. The stem age for Freziera was estimated to be 13.786 Ma (95% highest probability density [HPD]=40.733-29.676 Ma) and the crown age 12.648 Ma (95% HPD=17.227-10.870 Ma) (Fig. 4; Supplementary Fig. 3). The two daughter lineages of the crown node--the Humiriifolia clade and core Freziera--each have long branches leading to their respective crown radiations at 5.491 Ma (95% HPD=6.717-3.623 Ma) and 6.9164 Ma (95% HPD=7.771-5.481 Ma; Fig. 4; Supplementary Fig. 3)

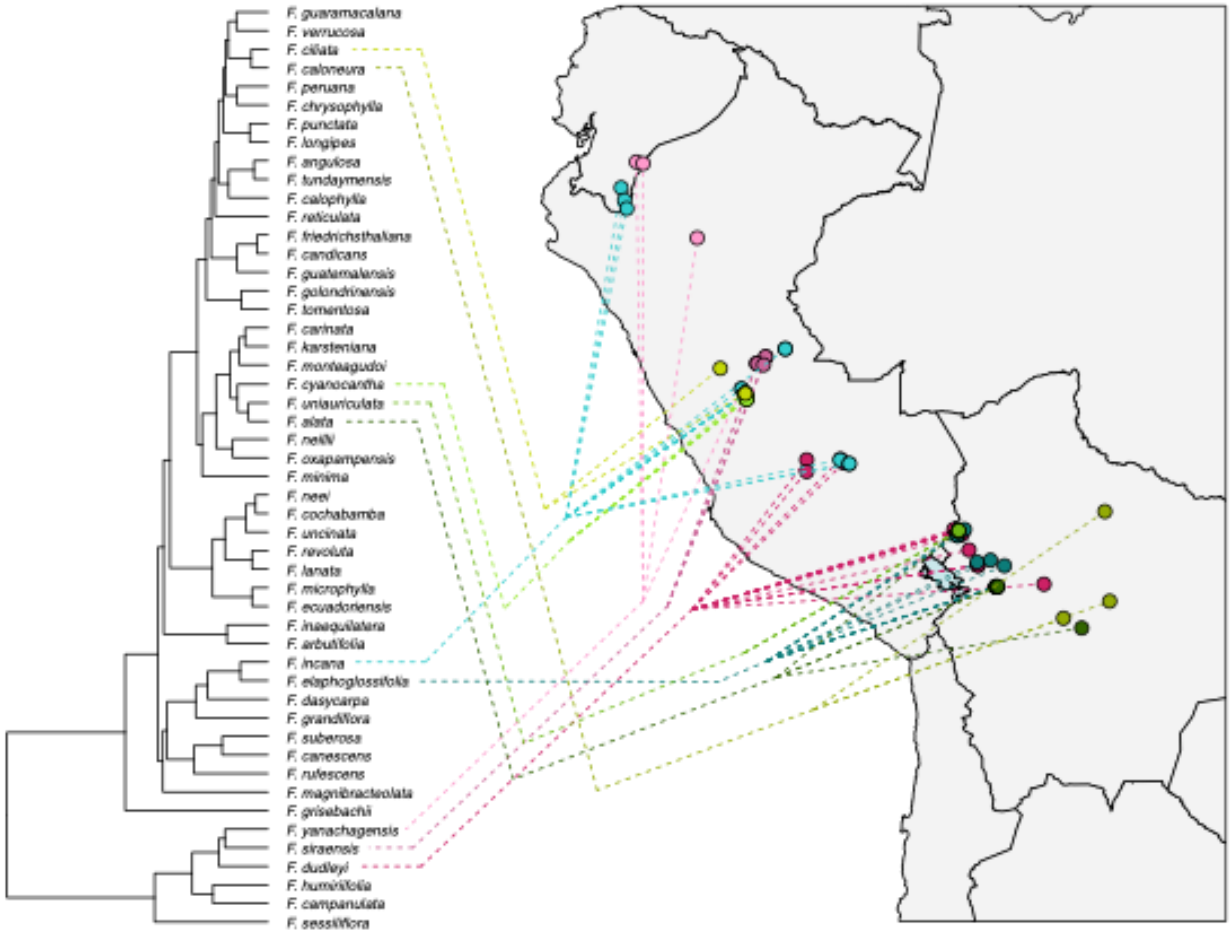
FIGURE 4.


Mesoamerica
Guiana shield
northern Andes
central Andes
F. guaramacalana
F. verrucosa
F. ciliata
F. caloneura
F. peruana
F. chrysophylla
F. punctata
F. longipes
F. angulosa
F. tundaymensis
F. calophylla
F. reticulata
F. friedrichsthaliana
F. candicans
F. guatemalensis
F. golondrinensis
F. tomentosa
F. carinata
F. karsteniana
F. monteagudoi
F. uniauriculata
F. alata
F. cyanocantha
F. neillii
F. oxapampensis
F. minima
F. neei
F. cochabamba
F. uncinata
F. lanata
F. revoluta
F. microphylla
F. ecuadoriensis
F. inaequilatera
F. arbutifolia
F. elaphoglossifolia
F. incana
F. dasycarpa
F. grandiflora
F. suberosa
F. canescens
F. rufescens
F. magnibracteolata
F. grisebachii
F. yanachagensis
F. siraensis
F. dudleyi
F. humiriifolia
F. campanulata
F. sessiliflora
My
12
10
8
6
4
2
0
Miocene
Pliocene
Pleist.
Number of lineages
50
40
30
20
10
0
Elevation (m)
5000
4000
3000
2000
1000
0

Biogeographic Reconstruction

The northern Andes are the most probable ancestral distribution for *Freziera*. The northern Andes were estimated as the ancestral area for the MRCA of the genus, nodes along the backbone of the phylogeny, and for MRCAs for each of the named clades (Fig. 4). From the northern Andes, repeated dispersal into all other geographic areas defined occurred. The central Andes were the main sink for northern Andean dispersal; ten dispersals from the northern Andes to the central Andes and one range expansion into the central Andes (*F. reticulata*) were estimated. Dispersals into the Central Andes often resulted in small clades of species that are currently geographically isolated (e.g., *F. yanachagensis* in northern and central Peru, *F. siraensis* in central Peru and *F. dudleyi* in southern Peru/western Bolivia; *F. incana* (Peru) and *F. elaphoglossifolia* (Bolivia); *F. cyanocantha* (Peru) and *F. alata* plus *F. uniauriculata* (Bolivia); and *F. ciliata* (Peru) and *F. caloneura* (Bolivia); Fig. 5). No dispersal from the central Andes to the northern Andes was inferred, through three range expansions back into the northern Andes were reconstructed (*F. chrysophylla*, the ancestor of *F. karsteniana*, and *F. yanachagensis*). From a shared widespread Andean ancestor with *F. karsteniana*, *F. carinata* retained a northern Andean distribution and expanded into the Guiana Shield. A separate dispersal from the northern Andes to the Guiana Shield was estimated for *F. guaramacalana*. Finally, three separate expansions into Mesoamerica from the northern Andes were inferred (*F. calophylla*; the ancestor of *F. candicans*, *F. friedrichsthaliana*, and *F. guatemalensis*; and *F. grisebachii*). *Freziera friedrichsthaliana* and *F. guatemalensis* were estimated to have become restricted to Mesoamerica from a widespread northern Andean-Mesoamerican ancestor.

FIGURE 5.



Climate and Soil Data

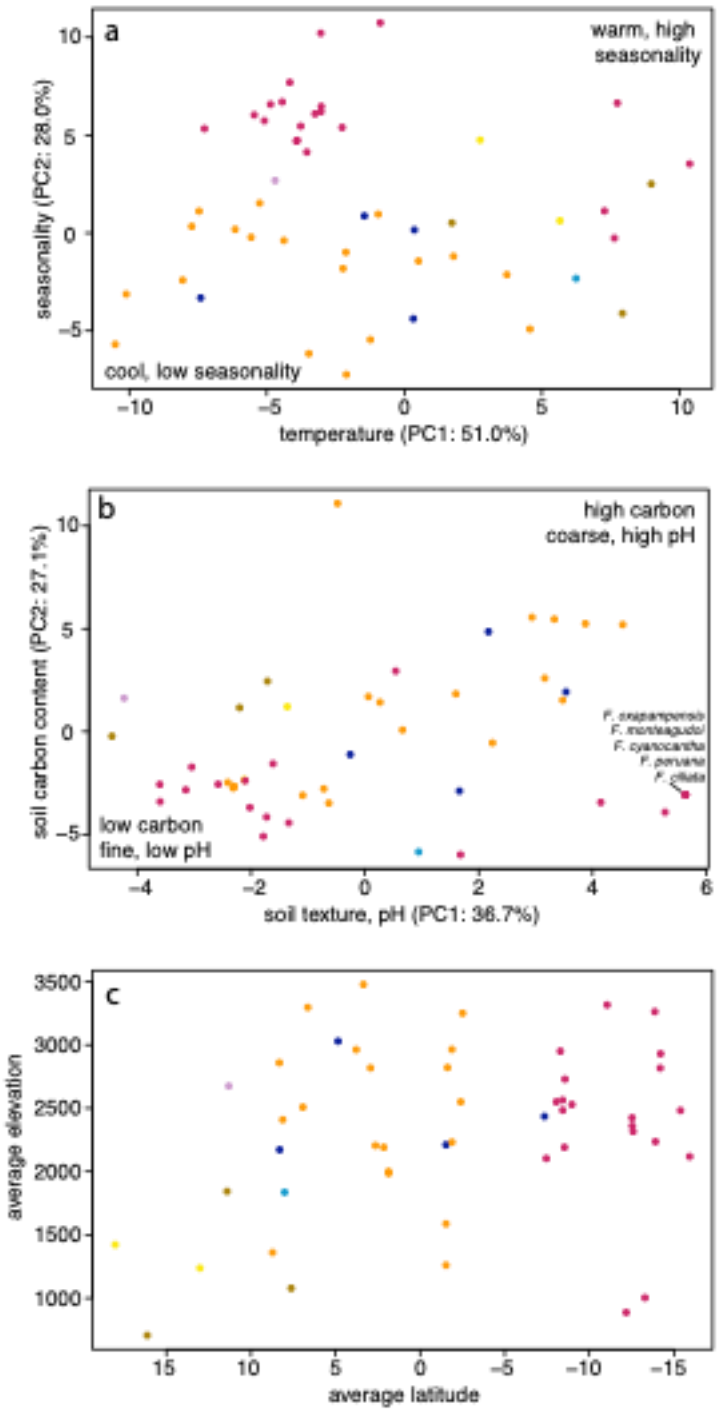
Climate and soil data were extracted based on 1,204 georeferenced specimens of *Freziera*. Occurrence points per species ranged from 1 (*F. monteagudoi* and *F. tundaymensis*) to 154 (*F. candicans*), with an average of 26 occurrences per species. Per species climate and soil averages can be found in Supplementary Table 3.

Phylogenetic PCA

The first two principal components of climate explained 79.0% of the variance in the data (PC1=51.0% and PC2=28.0%). Variables contributing to climate PC1 included elevation, annual mean temperature (Bio 1), mean temperature of the warmest quarter (Bio10), mean temperature

of wettest quarter (Bio 8), mean temperature of driest quarter (Bio 9), mean temperature of coldest quarter (Bio 11), minimum temperature of coldest month (Bio 6), maximum temperature of warmest month (Bio 5), precipitation of wettest quarter (Bio 16), precipitation of wettest month (Bio 13), annual precipitation (Bio 12), precipitation of warmest quarter (Bio 18). PC1 was thus an appropriate proxy for temperature, separating species from warmer, low-elevation environments from those from cooler, high elevation habitats. Climate PC2 included temperature annual range (Bio 7), precipitation of coldest quarter (Bio 19), precipitation of driest quarter (Bio 17), precipitation of driest month (Bio 14), precipitation seasonality (Bio 15), mean diurnal range (Bio 2), temperature seasonality (Bio 4), Isothermality (Bio 3). PC2 was an appropriate proxy for seasonality, separating habitats with higher seasonality, especially precipitation seasonality, from those with lower seasonality by PC2 (Fig. 6).

FIGURE 6.



The first two soil PCs explained 63.8% of the variance in the data; PC1 and PC2 explained 36.7% and 27.1% of the variance, respectively. Topsoil sand fraction by percent weight, topsoil reference bulk density, available water storage capacity, topsoil silt fraction by

percent weight, topsoil clay fraction by percent weight, topsoil pH in water, and topsoil bulk density contributed to PC1 ("soil texture, pH"). Topsoil organic carbon by percent weight, dominant soil type, topsoil carbon content, and area weighted topsoil carbon content contributed to PC2 ("soil carbon content"). The first PC separated coarser soils with lower pH from more fine-textured soils with higher pH. The second PC separated soils with high carbon content from low carbon soils. Principal component scores for the first two PCs for climat and soil data are listed in Supplementary Table 3.

DISCUSSION

Freziera is an understudied genus, but one that provides a good system to understand extrinsic factors that influence Andean cloud forest diversification. Biological factors—like short divergence times, high diversification rates, complex evolutionary histories, and polyploidy—and practical factors—like reliance on herbarium specimens for tissue resulting in poor-quality input DNA and few genomic resources available for probe design—are common barriers to phylogenetic inference in Andean groups and, therefore, rigorous investigation of evolutionary hypotheses. We produce the first phylogenetic hypothesis for *Freziera* using almost entirely herbarium tissue as the source of DNA for hybrid-enriched target sequence capture using the Angiosperms353 universal bait set. This dataset highlights the many challenges of working with understudied clades even in the genomic era: no a priori phylogenetic hypothesis, poor-quality DNA (short contigs and low coverage in non-coding regions), and relatively low probe specificity at the majority of targeted sites. We examine the effect of filtering genes and gene trees by different empirical criteria on species tree inference and show that inaccurate phylogenetic signal in gene trees was a greater problem in our dataset than low phylogenetic

signal, the common metric by which gene trees have been selected in genomic datasets. The removal of noise improved concordance and support. However, summary methods appear robust to such noise with a sufficient number of loci (~300 vs. ~150 in our datasets; Table 2), and, even with improved concordance in analyses accounting for inaccurate signal, conflict among gene trees remained high (Table 2; Supplementary Fig. 2). This is similar to results from other studies, suggesting that discordance may be a rule among Andean plant radiations (Morales-Briones, Liston, and Tank 2018; Vargas, Ortiz, and Simpson 2017; Bagley et al. 2020; Meerow, Gardner, and Nakamura 2020).

With the first phylogeny of *Freziera*, we reconstruct the biogeographic history of this widespread, Neotropical cloud forest genus. Focused biogeographic studies in cloud forest genera are few, as the often showy and diverse floral morphologies and/or frequent shifts in life history traits tend to be the primary focus of evolutionary studies. Without these confounding factors in *Freziera*, we identify the roles of environmental adaptation and dispersal in diversification. This is one of the first biogeographic studies in cloud forest plants to identify the Northern Andes as the ancestral distribution and source region and to distinguish patterns between the northern and central Andes. Given that the northern Andes typically house more taxonomic diversity than the central Andes for Andean-centered lineages (Gentry, 1982), our results highlight the need for more-detailed Andean cloud forest biogeographic studies to better understand the role of the northern Andes in generating biodiversity (i.e., is it a species pump or a sink that sparks diversification?).

Phylogenetics of Andean Radiations

One of the barriers to the study of Neotropical diversification is the difficulty resolving phylogenies of recent, rapid Andean radiations. Despite *Freziera*'s seemingly tractable evolutionary history—75 spp. with an estimated crown age of 12.6 Ma compared to 500+ spp. in ~5 Ma for the centropogonid clade of Lobelioideae (Lagomarsino et al. 2014, 2016), 200 spp. in 3.5 Ma for core *Puya (Givnish et al. 2011; Jabaily and Sytsma 2012)*, and 81 spp. in <2 Ma for Andean *Lupinus (Hughes and Eastwood 2006)*—we find similar hallmarks of explosive radiations in our phylogenetic results. High conflict among gene tree topologies (Supplementary Fig. 2) was found to underlie a relatively stable species tree for *Freziera* (Fig. 3c). Gene tree heterogeneity may be due to high levels of ILS, which is common in Andean radiations (Gómez-Gutiérrez et al. 2017; Pouchon et al. 2018; Nicola et al. 2019). However, a mixture of long and short internal branches, measured both in millions of years and coalescent units (Fig. 3), suggest only moderate ILS in *Freziera*. Limitations of using degraded DNA from herbarium specimens (e.g., low-quality data, poor/differential amplification, and low coverage) as well as a universal bait set within a genus (e.g., potential for low phylogenetic signal among close relatives for highly conserved genes) point to GTEE as another source of disagreement among gene trees and a source of error in species tree estimation. We explore how species tree inference with summary methods is impacted by data filtering with different empirical criteria for identifying GTEE, and build on the nascent body of literature evaluating the phylogenetic utility of paralogous loci.

Gene tree filtering.—Without an explicit method to estimate GTEE in empirical systems, several implicit methods have been used to identify genes with low phylogenetic signal. Alignment length, the number/proportion of variable or parsimony informative sites, total tree length, the proportion of internal branch lengths, and average bootstrap support of gene trees

have all been used as metrics by which to filter gene trees when GTEE can not be quantified (Leaché et al. 2014; Liu et al. 2015; Shen et al. 2016; Blom et al. 2017). However, when applied to the *Freziera* dataset we find that all but *1000bp* resulted in worse species trees relative to the *all.orthologs* dataset as defined by various metrics: higher discordance, measured by final normalized quartet score; lower support, both in average node support and proportion of well-supported nodes; and greater RF distance from the consensus topology (Table 2). Filtering for alignment length (i.e., the *1000bp* dataset) resulted in 300 of the 314 orthologs, so its similar performance to *all.orthologs* is perhaps not surprising. The higher performance of all loci relative to what should be the most informative loci after filtering for low phylogenetic signal suggests that strong but inaccurate signal is a larger problem than lack of signal in our dataset. Indeed, with an average alignment length over 3,000 bp and an average of almost 5 parsimony informative sites per internal branch in the tree for the ingroup, phylogenetic resolution would not seem to be the primary issue in this dataset as it may be for shorter targets, like UCEs (Meiklejohn et al., 2016) or RADseq data (Eaton et al. 2017).

Hidden paralogs, biological or artifactual, are not typically accounted for in phylogenomic datasets (Smith and Hahn, 2021a; 2021b), but they are a potential source of inaccurate signal and additional conflict in datasets for which gene tree discordance is already high. This is especially true in herbariomic datasets and systems for which little is known about the history of gene or genome duplication in the system, a common limitation in plant groups (Li and Barker 2020). Ploidy levels in *Freziera* are unknown due to few genomic resources (one transcriptome from *Ternstroemia gymnanthera* [Carpenter et al. 2019; One Thousand Plant Transcriptomes Initiative 2019] and one shotgun genome from *Anneslea fragrans* [Sun et al. 2017]) and fewer than 10 chromosome counts available for Pentaphylacaceae. There has been at

least one whole genome duplication (WGD) event in an ancestor of Pentaphylacaeae at a deep node in the inclusive order Ericales (Larson et al. 2020), and available chromosome counts (i.e., *Adinandra* [2 spp., n=42], *Cleyera* [1 spp., n=45], *Eurya* [4 spp., n=21, 29, 42], and *Ternstroemia* [2 spp., n=20, 25]; from Chromosome Counts Database [CCDB;(Rice et al. 2015); ccdb.tau.ac.il]) suggest either more recent duplication events or loss of copies across taxa. Given that (1) there is an ancient history of WGD and the suggestion of different ploidy levels in Pentaphylacaceae, (2) the dataset included known paralogs, (3) gene tree-species tree discordance was high, and (4) we had a high proportion of herbariomic data, we expected that cryptic paralogs are an important factor influencing discordance.

Filtering by bipartition support, the criterion which most improved species tree inference, identified gene trees with topologies consistent with cryptic paralogs. In our dataset, these gene trees typically exhibited polyphyly of individuals representing monophyletic species and deep divergences between clades, including members of those spuriously polyphyletic species (Figure 2). This pattern points to a largely artifactual source of cryptic paralogs, as repeated differential loss within species seems an unlikely biological phenomenon. As troubling as the prospect of unfiltered paralogs may be for empiricists, these artifactual loci may be easier to detect than biological pseudo-orthologs (Smith and Hahn 2021b), because the aforementioned pattern in gene trees is so striking. Confident identification of biological pseudo-orthologs would likely require better genomic resources than are available for most plant systems. These remain a potential source of conflict hidden in our dataset and others, but one to which methods that model ILS, like ASTRAL, may be sufficiently robust (Smith and Hahn 2021b). Our results suggest that these methods are also reasonably robust to the noise introduced by errant phylogenetic signal--in our case hypothesized to be due to cryptic paralogs; however, this may

not hold in lineages with higher rates of ILS. While both *bipartition* and *clocklike.bipartition* did improve gene tree concordance and support, discordance remained high and gains in support were minimal (Table 2).

The greatest benefit of filtering was topological, not improvement in concordance or support. The *all.orthologs* species tree, generated with unfiltered loci, did not recover all major clades identified in the consensus topology; in particular, the Elaphoglossifolia group was not monophyletic. Filtering by bipartition support resulted in a species tree consistent with the consensus tree, including monophyly of the nine named clades and their relative branching order in the consensus tree. Further filtering by bipartition support and then by root-to-tip variance also resulted in a monophyletic Elaphoglossifolia group, but lost full resolution of the two clades within the Candicans group. These datasets highlight the trade-offs of more data versus more-curated data. The reduction of noise through further filtering improved resolution and support in some areas of the tree (Supplementary Fig. 3b,c), but the loss of information by using fewer loci also resulted in lowered support and resolution in others (Supplementary Fig. 1b,i). The *1000bp* dataset, which had a high degree of overlap with *all.orthologs*, also recovered a monophyletic Elaphoglossifolia group. This suggests that a few uninformative loci—and generally weak resolution of that relationship—could have prevented resolution of major relationships in the *all.orthologs* dataset. The *1000bp* dataset was equally similar to the consensus topology as the species tree for *bipartition*. Using alignment length (i.e., greater than 1000 bp in our study) as a proxy for the amount of phylogenetic information in loci may be helpful for resolving major relationships in phylogenomic studies. However, filtering by the hypothetical accuracy of phylogenetic information via bipartition support accomplished the same goal and resulted in

higher gene tree concordance and support for species trees—a desirable goal for phylogenomic analysis.

Ours is representative of many target sequence capture datasets: we included a high proportion of data collected from museum specimens and relied on a universal probe set due to lack of publicly available genomic data for our focal group. Recapitulating filtering procedures from previous empirical studies (i.e., selecting for the greatest potential phylogenetic informativeness) would have resulted in an inaccurate species tree and misleading evolutionary inferences in almost all cases. However, these criteria were developed for datasets with different challenges and limitations, and will thus likely perform well for many research questions. We recommend that empirical phylogeneticists, especially those relying on museum specimens and/or universal probe sets, evaluate various cleaning and filtering techniques and to tailor data curation based on the unique properties of their individual datasets. In our case, inaccurate phylogenetic signal among gene trees consistent with cryptic paralogy was a major source of GTEE that had a negative impact on gene tree concordance and support. Removing this spurious signal by filtering by bipartition support resulted in a better species tree— despite including only roughly half the number of loci (i.e., 314 vs. 166 loci in ASTRAL-III analyses without known paralogs, and 331 vs. 197 loci in ASTRAL-Pro analyses with known paralogs). The presence of cryptic paralogs and their influence on phylogenetic inference is worth investigating in similar datasets, particularly those that aim to resolve shallow phylogenetic relationships using museomic data within clades with a known history of genome duplication.

Inclusion of paralogs.—The recent surge in genomic datasets and the ubiquity of genes that have undergone gene duplication and loss in those datasets has prompted the expansion of

phylogenetic methods to accommodate GDL. However, the utility of multi-copy genes for phylogenetic inference is only beginning to be explored. Gardner et al. (2020) found that the inclusion of paralogs increased disagreement among trees inferred by different methods from coding regions only, but reduced disagreement when both coding and non-coding regions were included (i.e., "supercontigs" from HybPiper). We similarly found that the addition of known paralogs to our datasets, which included non-coding regions, generally improved topological agreement. Species trees inferred from datasets including paralogs were more similar both to each other and to the consensus tree than those from orthologous loci alone. In several cases, the addition of paralogs to the dataset recovered major relationships in the consensus topology where the orthologs alone could not: both the placement of *F. minima* in the *clocklike.bipartition* tree and the non-monophyly of the Elaphoglossifolia group in the *all.orthologs* tree were resolved by the addition of paralogs. However, the inclusion of paralogs reduced concordance in all cases and support in most, a result also found by Gardner et al. (2020). It seems that paralogs are helpful for topological resolution, especially at deep nodes, though this is at the expense of gene tree concordance and overall support.

The ability to include loci that have a history of GDL is desirable, especially in systems that have a history of genome duplication, as additional copies may make up a significant portion of the data collected (Johnson et al., 2016; Morales-Briones et al., 2020). However, it is important that empirical systematists are mindful of the potential trade-off between support and resolution in the context of their studies. Large, phylogenomic datasets carry the promise of additional data that can help resolve difficult relationships, whether shallow relationships in a rapid radiation or deep nodes along the backbone. The inclusion of multi-copy genes further increases the amount of data available for phylogenetic inference, with early evidence from this

study and Gardner et al. (2020) suggesting these data best serve the goal of resolving deep nodes. While early empirical studies have found this utility for paralogs in phylogenetic inference, we have also found that the resolution provided by paralogs comes at a cost: lower support. Just as researchers should be wary of high bootstrap values from concatenation-based methods despite underlying discordance among genes (Kubatko and Degnan 2007; Degnan and Rosenberg 2009; Sayyari and Mirarab 2016), we should perhaps be careful not to overinterpret higher support in ortholog-only trees. That said, a weakness of empirical studies is the lack of knowledge of the true species tree. With the expansion of methods that incorporate GDL in their models and the emergent trends in empirical systems, studies using simulations from a known tree would be beneficial for better understanding the observed trade-offs between backbone resolution and support.

Distinct Modes of Diversification Across Regions within the Montane Cloud Forest Biome

The Mid- to Late Miocene crown age of *Freziera* adds to a growing body of literature pointing towards shared Miocene origins across Andean cloud forest plant groups (Lagomarsino et al. 2016; Givnish et al. 2014; Spriggs et al. 2015; Schwery et al. 2015; Testo et al. 2019). As in other lineages, the expansion of this biome in the northern Andes following orogenic events in the Neogene seems to be particularly important in spurring diversification in *Freziera* (Fig. 4). The northern Andes are an important area in the early diversification of the genus: they are the inferred ancestral distribution of the genus, its two principal radiations (the Humiriifolia clade and core *Freziera*), and the majority of backbone nodes. The two principal radiations both have relatively long stem lineages followed by Late Miocene diversification leading to extant diversity. This timing suggests radiation in these clades follows major mountain building events

in the northern Andes ca.10-7 Ma and coincides with the final uplift of the northern Andes in the Pliocene (Fig. 4; Gregory-Wodzicki, 2000; Montgomery et al. 2001; Graham, 2009; Hoorn et al., 2010; Armijo et al. 2015).

The northern Andes also act as a source region from which dispersal into other montane Neotropical regions occurs in *Freziera*. This contrasts with the most common standing hypothesis that the older and higher central Andes, which could have supported cloud forest communities earlier than the northern Andes, are a source area for cloud forest clades, and that the northern Andes, with its three cordilleras that provide a greater heterogeneity of microclimates and additional opportunities for vicariance relative to the single cordillera of the Central Andes, are a sink that ignites diversification (Simpson, 1983). We find support for the opposite scenario in *Freziera*: ancestral diversification in the northern Andes is punctuated by twelve dispersals into other areas, with the central Andes acting as the most frequent sink for northern Andean exports. In total, we infer nine movements into the central Andes, two into Central America, and one to the Guiana Shield. Both small radiations (including a Central American subclade of the Calophylla clade and the majority of species in the Lanata clade) and single speciation events result from movement to newly colonized regions. While dispersal out of the northern Andes has been frequent in *Freziera*, movement back into this ancestral range is uncommon. The only estimated mode of dispersal into the northern Andes is through range expansion; species maintain distributions in other areas as well. Few biogeographic studies identify northern Andes as a source region for plants (e.g., Cinchoneae (Antonelli et al. 2009) and Neotropical *Hedyosmum* (Antonelli and Sanmartín 2011), despite the fact that the northern Andes tend to house greater taxonomic diversity of Andean-centered lineages (Gentry, 1982). However, biogeographic analyses of plant clades with primarily cloud forest distributions are

underrepresented in the literature in proportion to the species-richness of this biome (Hughes et al. 2012) and fewer still examine the northern and central Andeans as separate regions. The lack of evidence for the northern Andes as a source area in plant lineages may simply be a bias in the literature and represent an area where additional research would be particularly fruitful.

Northern Andean species of Freziera exhibit more variation in the environmental niches that they occupy relative to Central Andean species, resulting in a more even distribution in both climate and soil space than Central American species (Fig. 6). It is thus likely the ancestral northern Andean radiation in *Freziera* may have been spurred by uplift-driven vicariance followed by local climatic adaptation to the many heterogeneous and dynamic mid-montane mesic habitats that originated in the northern Andes during active mountain uplift. This putative ancestral diversity of climate and soil preferences would facilitate the northern Andes as a source region to other areas, with pre-adapted lineages filtering into Central American and Central Andean habitats via dispersal, while retaining diversity in the northern Andes. Further, global vegetation reconstructions of the late Miocene (11.61-7.25 Ma) show that tropical evergreen forests receded to the area around the northern Andes during this period, whereas the central Andean region was dominated by savannas and deciduous forest (though recent paleobotanical studies reconstruct a wetter Miocene than previously suggested for the Neogene flora in the Central Andean Plateau; Martinez et al., 2020). If central Andean climatic conditions were unsuitable for *Freziera* at that time, this would fit with the observed pattern of northern Andean diversification at the onset of mountain uplift followed by dispersal to the central Andes as modern montane forest analogs emerged.

While the northern Andes clearly act as a source of diversity for *Freziera*, the central Andes house nearly equal taxonomic diversity (ca. 24 spp.) as the northern Andes (27 spp., plus

ca. 10 occurring in both regions; Santamaría-Aguilar and Monro, 2019). While northern Andean species cover a broad range of climate regimes, consistent with this region being one of the most climatically variable regions globally (Rahbek et al. 2019), most central Andean species are clustered either in cooler, more-seasonal areas or warmer, more-seasonal climates (Fig. 6a). Climatic similarity between central Andean communities likely facilitated establishment of new species following dispersal or biome fragmentation between them. Supporting this, closely related central Andean species are often distributed in different regions of the central Andes (Fig. 5). *Freziera* reveals different, but complementary patterns for Andean cloud forest diversification driven primarily by extrinsic factors: micro-scale allopatry and niche differentiation are evolutionary themes in the northern Andes, whereas macro-scale allopatry and niche conservatism are common in the central Andes.

CONCLUSION

We present the first phylogenetic hypothesis and macroevolutionary study of *Freziera.* Our molecular and geospatial data are largely gathered from herbarium specimens, highlighting the role of collections data in macroevolutionary research (Lendemer et al. 2020). Our results suggest that diversification dynamics differ between the northern and central Andean regions, as has been documented in other lineages (Pérez-Escobar et al. 2017). While the actively rising northern Andeans provided the backdrop along which *Freziera* diversified and filled the majority of its currently realized niche space, dispersal into the Central Andes was associated with in situ radiations within similar, pre-existing habitats. This demonstrates that landscape-scale heterogeneity in topography and climate have profound impacts on evolution within lineages, and may help explain the extraordinary species richness of Andean cloud forests— even in the

absence of complex ecological interactions or labile morphological evolution that are landmarks of many rapid Andean radiations.

We further explored the need to carefully examine and filter data in empirical phylogenomic datasets. This is increasingly acknowledged as a necessary step in plant phylogenomics due to complexities of plant genome evolution, including frequent genome duplications (Landis et al. 2018; Ren et al. 2018). Via these filtering steps, we found cryptic paralogs in our dataset, which we were able to remove using bipartition support criteria. Removing loci with aberrant phylogenetic signal decreased gene tree discordance in the final dataset and improved phylogenetic support. In addition, approximately 9% of our dataset was represented in multi-copy paralogous loci. Incorporating these paralogs using a method developed to analyze multi-copy loci improved resolution of deep relationships in *Freziera*, even while overall support was lower. This tradeoff between resolution and support resulting from the incorporation of paralogs is an important consideration for empirical and theoretical phylogeneticists alike. Given the numerous biological and practical factors that contribute to complexities in phylogenomic datasets, various data processing techniques should be investigated to determine the extent to which more vs. more-curated data is appropriate for specific phylogenetic questions.

Finally, our results further have major implications for understanding the origin of plant biodiversity in the world's most species-rich biome: Andean cloud forests. Despite its global importance, evolutionary history of taxa in this region remains understudied—in large part due to the difficulty in inferring well-supported phylogenies for its often species-rich radiations. We encountered some of these difficulties in our phylogenomic analyses: short branch lengths characteristic of Andean plant radiations obscured relationships within some subclades and some deep nodes could not be resolved with high support, perhaps suggesting a deep history of

introgression. Further, our reliance on herbarium tissue may have increased in the proportion of artifactual cryptic paralogs in our dataset. Despite these challenges, many relationships within and among clades of *Freziera* were supported across analyses, and support among these relationships increased when filtering methods maximized our chances of removing highly discordant, cryptically paralogous loci. Due to the young age of the Andes and their resident clades, the dynamic and continental-scale landscape change, and species richness of many endemic clades, Andean plant clades will likely remain some of the most challenging to resolve even as phylogenomic datasets increase in size and analytical methods improve.

ACKNOWLEDGEMENTS

This research was funded by a Louisiana Board of Regents Research Competitiveness Subprogram grant and by the LSU College of Science and Office of Research and Economic Development. We would like to thank the Missouri Botanical Garden (MO) for their access to their important collections. We thank Brant Faircloth, Matthew Johnson, Carl Oliveros, and Jessie Salter for their guidance in library preparation, and Brant Faircloth for access to laboratory equipment. Computational analyses were performed on LSU High Performance Computing's SuperMike cluster. This manuscript benefited from feedback from Laymon Ball, Janet Mansaray, and Diego Paredes-Burneo, while the taxonomic expertise of Daniel Santamaría-Aguilar benefitted us throughout the design and implementation of this research.

SUPPLEMENTARY MATERIAL

Data available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.jsxksn09g

REFERENCES

Anisimova M., Gil M., Dufayard J.-F., Dessimoz C., Gascuel O. 2011. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. Syst. Biol. 60:685–699.

Antonelli A., Kissling W.D., Flantua S.G.A., Bermúdez M.A., Mulch A., Muellner-Riehl A.N., Kreft H., Linder H.P., Badgley C., Fjeldså J., Others. 2018. Geological and climatic influences on mountain biodiversity. Nat. Geosci. 11:718–725.

Antonelli A., Sanmartín I. 2011. Why are there so many plant species in the Neotropics? Taxon. 60:403–414.

Bakker F.T. 2017. Herbarium genomics: skimming and plastomics from archival specimens. Webbia. 72:35–45.

Bakker F.T., Lei D., Yu J., Mohammadin S., Wei Z., van de Kerke S., Gravendeel B., Nieuwenhuis M., Staats M., Alquezar-Planas D.E., Holmer R. 2015. Herbarium genomics: plastome sequence assembly from a range of herbarium specimens using an Iterative Organelle Genome Assembly pipeline. Biol. J. Linn. Soc. Lond. 117:33–43.

Beaulieu J.M., O'Meara B.C. 2018. Can we build it? Yes we can, but should we use it? Assessing the quality and value of a very large phylogeny of campanulid angiosperms. Am. J. Bot. 105:417–432.

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

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

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

Borowiec M.L. 2016. AMAS: a fast tool for alignment manipulation and computing of summary statistics. PeerJ. 4:e1660.

Braun G., Mutke J., Reder A., Barthlott W. 2002. Biotope patterns, phytodiversity and forestline in the Andes, based on GIS and remote sensing data. In: Körner C., Spehn E.M., editors. Mountain Biodiversity: A Global Assessment. London, UK: Parthenon Publishing. p. 75–89.

Breinholt J.W., Carey S.B., Tiley G.P., Davis E.C., Endara L., McDaniel S.F., Neves L.G., Sessa E.B., von Konrat M., Chantanaorrapint S., Fawcett S., Ickert-Bond S.M., Labiak P.H., Larraín J., Lehnert M., Lewis L.R., Nagalingum N.S., Patel N., Rensing S.A., Testo W., Vasco A., Villarreal J.C., Williams E.W., Burleigh J.G. 2021. A target enrichment probe set for resolving the flagellate land plant tree of life. Appl. Plant Sci. 9:e11406.

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

Brown J.W., Walker J.F., Smith S.A. 2017. Phyx: phylogenetic tools for unix. Bioinformatics. 33:1886–1888.

Carpenter E.J., Matasci N., Ayyampalayam S., Wu S., Sun J., Yu J., Jimenez Vieira F.R., Bowler C., Dorrell R.G., Gitzendanner M.A., Li L., Du W., K Ullrich K., Wickett N.J., Barkmann T.J., Barker M.S., Leebens-Mack J.H., Wong G.K.-S. 2019. Access to RNA-sequencing data from 1,173 plant species: The 1000 Plant transcriptomes initiative (1KP). Gigascience. 8:giz126.

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

Copetti D., Búrquez A., Bustamante E., Charboneau J.L.M., Childs K.L., Eguiarte L.E., Lee S., Liu T.L., McMahon M.M., Whiteman N.K., Wing R.A., Wojciechowski M.F., Sanderson M.J. 2017. Extensive gene tree discordance and hemiplasy shaped the genomes of North American columnar cacti. Proc. Natl. Acad. Sci. U. S. A. 114:12003–12008.

Cosentino S., Iwasaki W. 2019. SonicParanoid: fast, accurate and easy orthology inference. Bioinformatics. 35:149–151.

Cuello N.L., Santamaría-Aguilar D. 2015. A New Species of *Freziera* (Pentaphylacaceae) from the Venezuelan Andes. hpib. 20:147–150.

Davis S.D., Heywood V.H., Herrera-MacBryde O., Villa-Lobos J., Hamilton A.C. 1997. Centres of plant diversity: a guide and strategy for their conservation. Volume 3. The Americas. The

Worldwide Fund for Nature (WWF)/The World Conservation Union (IUCN).

Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.

Donoghue M.J., Edwards E.J. 2019. Model clades are vital for comparative biology, and ascertainment bias is not a problem in practice: a response to Beaulieu and O'Meara (2018). Am. J. Bot. 106:327–330.

Donoghue M.J., Sanderson M.J. 2015. Confluence, synnovation, and depauperons in plant diversification. New Phytol. 207:260–274.

Drummond A.J., Ho S.Y.W., Phillips M.J., Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:e88.

Drummond A.J., Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7:214.

Drummond A.J., Suchard M.A., Xie D., Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29:1969–1973.

Eaton D.A.R., Spriggs E.L., Park B., Donoghue M.J. 2017. Misconceptions on missing Data in RAD-seq phylogenetics with a deep-scale example from flowering plants. Syst. Biol. 66:399–412.

Emms D.M., Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16:157.

Emms D.M., Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative

genomics. Genome Biol. 20:238.

Faircloth B.C. 2013. Illumiprocessor: a trimmomatic wrapper for parallel adapter and quality trimming. http://dx.doi.org/10.6079/J9ILL.

Faircloth B.C. 2016. PHYLUCE is a software package for the analysis of conserved genomic loci. Bioinformatics. 32:786–788.

Fick S.E., Hijmans R.J. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int. J. Climatol. 37:4302–4315.

Flantua S.G.A., O'Dea A., Onstein R.E., Giraldo C., Hooghiemstra H. 2019. The flickering connectivity system of the north Andean páramos. J. Biogeogr. 46:1808–1825.

Gardner E.M., Johnson M.G., Pereira J.T., Puad A.S.A., Arifiani D., Wickett N.J., Zerega N.J.C. 2020. Paralogs and off-target sequences improve phylogenetic resolution in a densely-sampled study of the breadfruit genus (*Artocarpus*, Moraceae). Syst. Biol. 70:558–575.

Gentry A.H. 1982. Neotropical floristic diversity: phytogeographical connections between Central and South America, Pleistocene climatic fluctuations, or an accident of the Andean orogeny? Ann. Mo. Bot. Gard. 69:557–593.

Gentry A.H., Dodson C.H. 1987. Diversity and biogeography of Neotropical vascular epiphytes. Ann. Mo. Bot. Gard. 74:205–233.

Givnish T.J. 2008. Comparative studies of leaf form: assessing the relative roles of selective pressures and phylogenetic constraints. New Phytol. 106:131–160.

Givnish T.J., Barfuss M.H.J., Van Ee B., Riina R., Schulte K., Horres R., Gonsiska P.A., Jabaily

R.S., Crayn D.M., Smith J.A.C., Winter K., Brown G.K., Evans T.M., Holst B.K., Luther H., Till W., Zizka G., Berry P.E., Sytsma K.J. 2011. Phylogeny, adaptive radiation, and historical biogeography in Bromeliaceae: insights from an eight-locus plastid phylogeny. Am. J. Bot. 98:872–895.

Givnish T.J., Barfuss M.H.J., Van Ee B., Riina R., Schulte K., Horres R., Gonsiska P.A., Jabaily R.S., Crayn D.M., Smith J.A.C., Winter K., Brown G.K., Evans T.M., Holst B.K., Luther H., Till W., Zizka G., Berry P.E., Sytsma K.J. 2014. Adaptive radiation, correlated and contingent evolution, and net species diversification in Bromeliaceae. Mol. Phylogenet. Evol. 71:55–78.

Givnish T.J., Spalink D., Ames M., Lyon S.P., Hunter S.J., Zuluaga A., Iles W.J.D., Clements M.A., Arroyo M.T.K., Leebens-Mack J., Endara L., Kriebel R., Neubig K.M., Whitten W.M., Williams N.H., Cameron K.M. 2015. Orchid phylogenomics and multiple drivers of their extraordinary diversification. Proc. Biol. Sci. 282:20151553.

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

Goodwin Z.A., Harris D.J., Filer D., Wood J.R.I., Scotland R.W. 2015. Widespread mistaken identity in tropical plant collections. Curr. Biol. 25:R1066–7.

Gravendeel B., Smithson A., Slik F.J.W., Schuiteman A. 2004. Epiphytism and pollinator specialization: drivers for orchid diversity? Phil. Trans. R. Soc. Lond. B Biol. Sci. 359:1523–1535.

Guindon S., Dufayard J.-F., Lefort V., Anisimova M., Hordijk W., Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59:307–321.

Guo Q., Kelt D.A., Sun Z., Liu H., Hu L., Ren H., Wen J. 2013. Global variation in elevational diversity patterns. Sci. Rep. 3:3007.

Hale H., Gardner E.M., Viruel J., Pokorny L., Johnson M.G. 2020. Strategies for reducing per-sample costs in target capture sequencing for phylogenomics and population genomics in plants. Appl. Plant Sci. 8:e11337.

Hart M.L., Forrest L.L., Nicholls J.A., Kidner C.A. 2016. Retrieval of hundreds of nuclear loci from herbarium specimens. Taxon. 65:1081–1092.

Hazzi N.A., Moreno J.S., Ortiz-Movliav C., 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.

Heled J., Drummond A.J. 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27:570–580.

Hijmans R.J., Van Etten J., Cheng J., Mattiuzzi M., Sumner M., Greenberg J.A., Lamigueiro O.P., Bevan A., Racine E.B., Shortridge A., Others. 2015. Package “raster.” R package.

Hoang D.T., Chernomor O., von Haeseler A., Minh B.Q., Vinh L.S. 2018. UFBoot2: improving the ultrafast bootstrap approximation. Mol. Biol. Evol. 35:518–522.

Hopkins M.J.G. 2007. Modelling the known and unknown plant biodiversity of the Amazon

Basin. J. Biogeogr. 34:1400–1411.

Hostettler S. 2002. Tropical montane cloud forests: a challenge for conservation. Bois et forets des Tropiques. 274:19–31.

Huang H., Knowles L.L. 2016. Unforeseen consequences of excluding missing data from next-generation sequences: simulation study of RAD sequences. Syst. Biol. 65:357–365.

Hughes C., Eastwood R. 2006. Island radiation on a continental scale: exceptional rates of plant diversification after uplift of the Andes. Proc. Natl. Acad. Sci. U. S. A. 103:10334–10339.

Humboldt A. von. 1808. Ansichten der Natur mit wiss. Erläuterungen. Tübingen: Cotta.

Humboldt A. von, Bonpland A. 1807. Essai súr la Géografie des Plantes. Paris, France: Chez Lavrault, Schoell.

Irisarri I., Meyer A. 2016. The Identification of the Closest Living Relative(s) of Tetrapods: Phylogenomic Lessons for Resolving Short Ancient Internodes. Syst. Biol. 65: 1057–1075.

Ivey C.T., DeSilva N. 2001. A test of the function of drip tips. Biotropica. 33:188–191.

Jabaily R.S., Sytsma K.J. 2012. Historical biogeography and life-history evolution of Andean *Puya* (Bromeliaceae). Bot. J. Linn. Soc. 171:201–224.

Jiao Y., Wickett N.J., Ayyampalayam S., Chanderbali A.S., Landherr L., Ralph P.E., Tomsho L.P., Hu Y., Liang H., Soltis P.S., Soltis D.E., Clifton S.W., Schlarbaum S.E., Schuster S.C., Ma H., Leebens-Mack J., dePamphilis C.W. 2011. Ancestral polyploidy in seed plants and angiosperms. Nature. 473:97–100.

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

Johnson M.G., Pokorny L., Dodsworth S., Botigué L.R., Cowan R.S., Devault A., Eiserhardt W.L., Epitawalage N., Forest F., Kim J.T., Leebens-Mack J.H., Leitch I.J., Maurin O., Soltis D.E., Soltis P.S., Wong G.K.-S., Baker W.J., Wickett N.J. 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.

Jørgensen P.M., Ulloa Ulloa C., León B., León-Yánez S., Beck S.G., Nee M., Zarucchi J.L., Celis M., Bernal R., Gradstein R. 2011. Regional patterns of vascular plant diversity and endemism. In: Herzog S.K., Martínez R., Jørgensen P.M., Tiessen H., editors. Climate Change and Biodiversity in the Tropical Andes. Inter-American Institute for Global Change Research (IAI) and Scientific Committee on Problems of the Environment (SCOPE). p. 192–203.

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

Katoh K., Standley D.M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30:772–780.

Kier G., Kreft H., Lee T.M., Jetz W., Ibisch P.L., Nowicki C., Mutke J., Barthlott W. 2009. A global assessment of endemism and species richness across island and mainland regions.

Proc. Natl. Acad. Sci. U. S. A. 106:9322–9327.

Kromer T., Kessler M., Robbert Gradstein S., Acebey A. 2005. Diversity patterns of vascular epiphytes along an elevational gradient in the Andes. J. Biogeogr. 32:1799–1809.

Kubatko L.S., Degnan J.H. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56:17–24.

Lagomarsino L.P., Antonelli A., Muchhala N., Timmermann A., Mathews S., Davis C.C. 2014. Phylogeny, classification, and fruit evolution of the species-rich Neotropical bellflowers (Campanulaceae: Lobelioideae). Am. J. Bot. 101:2097–2112.

Lagomarsino L.P., Condamine F.L., Antonelli A., Mulch A., Davis C.C. 2016. The abiotic and biotic drivers of rapid diversification in Andean bellflowers (Campanulaceae). New Phytol. 210:1430–1442.

Lagomarsino L.P., Frost L.A. 2020. The central role of taxonomy in the study of Neotropical biodiversity. Ann. Missouri Bot. Gard. 105:405–421.

Landis J.B., Soltis D.E., Li Z., Marx H.E., Barker M.S., Tank D.C., Soltis P.S. 2018. Impact of whole-genome duplication events on diversification rates in angiosperms. Am. J. Bot. 105:348–363.

Lanier H.C., Huang H., Knowles L.L. 2014. How low can you go? The effects of mutation rate on the accuracy of species-tree estimation. Mol. Phylogenet. Evol. 70:112–119.

Larson D.A., Walker J.F., Vargas O.M., Smith S.A. 2020. A consensus phylogenomic approach highlights paleopolyploid and rapid radiation in the history of Ericales. Am. J. Bot.

107:773–789.

Leaché A.D., Wagner P., Linkem C.W., Böhme W., Papenfuss T.J., Chong R.A., Lavin B.R., Bauer A.M., Nielsen S.V., Greenbaum E., Rödel M.-O., Schmitz A., LeBreton M., Ineich I., Chirio L., Ofori-Boateng C., Eniang E.A., Baha El Din S., Lemmon A.R., Burbrink F.T. 2014. A hybrid phylogenetic–phylogenomic approach for species tree estimation in African *Agama* lizards with applications to biogeography, character evolution, and diversification. Mol. Phylogenet. Evol. 79:215–230.

Legried B., Molloy E.K., Warnow T., Roch S. 2021. Polynomial-time statistical estimation of species trees under gene duplication and loss. J. Comput. Biol. 28:452–468.

Lendemer J., Thiers B., Monfils A.K., Zaspel J., Ellwood E.R., Bentley A., LeVan K., Bates J., Jennings D., Contreras D., Lagomarsino L., Mabee P., Ford L.S., Guralnick R., Gropp R.E., Revelez M., Cobb N., Seltmann K., Aime M.C. 2020. The Extended Specimen Network: a strategy to enhance US biodiversity collections, promote research and education. Bioscience. 70:23–30.

Liu L., Xi Z., Wu S., Davis C.C., Edwards S.V. 2015. Estimating phylogenetic trees from genome-scale data. Ann. N. Y. Acad. Sci. 1360:36–53.

Liu Y., Johnson M.G., Cox C.J., Medina R., Devos N., Vanderpoorten A., Hedenäs L., Bell N.E., Shevock J.R., Aguero B., Quandt D., Wickett N.J., Shaw A.J., Goffinet B. 2019. Resolution of the ordinal phylogeny of mosses using targeted exons from organellar and nuclear genomes. Nat. Commun. 10:1485.

Li Z., Barker M.S. 2020. Inferring putative ancient whole-genome duplications in the 1000

Plants (1KP) initiative: access to gene family phylogenies and age distributions. Gigascience. 9:giaa004.

Longo S.J., Faircloth B.C., Meyer A., Westneat M.W., Alfaro M.E., Wainwright P.C. 2017. Phylogenomic analysis of a rapid radiation of misfit fishes (Syngnathiformes) using ultraconserved elements. Mol. Phylogenet. Evol. 113:33–48.

Mai U., Mirarab S. 2018. TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. BMC Genomics. 19:272.

Markin A., Eulenstein O. 2020. Quartet-based inference methods are statistically consistent under the unified duplication-loss-coalescence model. arXiv: 2004.04299v1 [q-bio.PE].

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

Matzke N.J. 2013a. Probabilistic historical biogeography: new models for founder-event speciation, imperfect detection, and fossils allow improved accuracy and model-testing. Front. Biogeogr. 5:242–248.

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

Matzke N.J. 2014. Model selection in historical biogeography reveals that founder-event speciation is a crucial process in island clades. Syst. Biol. 63:951–970.

Mayrose I., Zhan S.H., Rothfels C.J., Magnuson-Ford K., Barker M.S., Rieseberg L.H., Otto S.P.

2011. Recently formed polyploid plants diversify at lower rates. Science. 333:1257.

McKain M.R., Johnson M.G., Uribe-Convers S., Eaton D., Yang Y. 2018. Practical considerations for plant phylogenomics. Appl. Plant Sci. 6:e1038.

Meiklejohn K.A., Faircloth B.C., Glenn T.C., Kimball R.T., Braun E.L. 2016. Analysis of a rapid evolutionary radiation using ultraconserved elements: evidence for a bias in some multispecies coalescent methods. Syst. Biol. 65:612–627.

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

Molloy E.K., Warnow T. 2018. To include or not to include: the impact of gene filtering on species tree estimation methods. Syst. Biol. 67:285–303.

Molloy E.K., Warnow T. 2020. FastMulRFS: fast and accurate species tree estimation under generic gene duplication and loss models. Bioinformatics. 36:i57–i65.

Morales-Briones D.F., Gehrke B., Huang C.-H., Liston A., Ma H., Marx H.E., Tank D.C., Yang Y. 2020. Analysis of paralogs in target enrichment data pinpoints multiple ancient polyploidy events in *Alchemilla* s.l. (Rosaceae). bioRxiv.:2020.08.21.261925.

Morales-Briones D.F., Kadereit G., Tefarikis D.T., Moore M.J., Smith S.A., Brockington S.F., Timoneda A., Yim W.C., Cushman J.C., Yang Y. 2021. Disentangling sources of gene tree discordance in phylogenomic data sets: testing ancient hybridizations in Amaranthaceae sl. Syst. Biol. 70:219–235.

Morales-Briones D.F., Liston A., Tank D.C. 2018a. Phylogenomic analyses reveal a deep history of hybridization and polyploidy in the Neotropical genus *Lachemilla* (Rosaceae). New Phytol. 218:1668–1684.

Morales-Briones D.F., Romoleroux K., Kolář F., Tank D.C. 2018b. Phylogeny and evolution of the Neotropical radiation of *Lachemilla* (Rosaceae): uncovering a history of reticulate evolution and implications for infrageneric classification. Syst. Bot. 43:17–34.

Muellner-Riehl A.N. 2019. Mountains as evolutionary arenas: patterns, emerging approaches, paradigm shifts, and their Implications for plant phylogeographic research in the Tibeto-Himalayan region. Front. Plant Sci. 10:195.

Muellner-Riehl A.N., Schnitzler J., Kissling W.D., Mosbrugger V., Rijsdijk K.F., Seijmonsbergen A.C., Versteegh H., Favre A. 2019. Origins of global mountain plant biodiversity: Testing the “mountain-geobiodiversity hypothesis.” J. Biogeogr. 46:2826–2838.

Mutke J., Barthlott W. 2005. Patterns of vascular plant diversity at continental to global scales. Biol. Skr. 55:521–531.

Mutke J., Weigend M. 2017. Mesoscale patterns of plant diversity in Andean South America based on combined checklist and GBIF data. Ber. d. Reinh.-Tüxen-Ges. 23:83–97.

Myers N., Mittermeier R.A., Mittermeier C.G., da Fonseca G.A., Kent J. 2000. Biodiversity hotspots for conservation priorities. Nature. 403:853–858.

Nguyen L.-T., Schmidt H.A., von Haeseler A., Minh B.Q. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol.

32:268–274.

Nicola M.V., Johnson L.A., 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.

Nute M., Chou J., Molloy E.K., Warnow T. 2018. The performance of coalescent-based species tree estimation methods under models of missing data. BMC Genomics. 19:286.

Ogilvie H.A., Bouckaert R.R., Drummond A.J. 2017. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. Mol. Biol. Evol. 34:2101–2114.

Oguchi R., Onoda Y., Terashima I., Tholen D. 2018. Leaf anatomy and function. In: Adams W.W. III, Terashima I., editors. The Leaf: A Platform for Performing Photosynthesis. Cham: Springer International Publishing. p. 97–139.

One Thousand Plant Transcriptomes Initiative. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. Nature. 574:679–685.

Paradis E., Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 35:526–528.

Parks K.E., Mulligan M. 2010. On the relationship between a resource based measure of geodiversity and broad scale biodiversity patterns. Biodivers. Conserv. 19:2751–2766.

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

Pérez-Escobar O.A., Chomicki G., Condamine F.L., Karremans A.P., Bogarín D., Matzke N.J., Silvestro D., Antonelli A. 2017. Recent origin and rapid speciation of Neotropical orchids in the world's richest plant biodiversity hotspot. New Phytol. 215:891–905.

Pouchon C., Fernández A., Nassar J.M., Boyer F., Aubert S., Lavergne S., Mavárez J. 2018. Phylogenomic analysis of the explosive adaptive radiation of the *Espeletia* Complex (Asteraceae) in the tropical Andes. Syst. Biol. 67:1041–1060.

Quintero I., Jetz W. 2018. Global elevational diversity and diversification of birds. Nature. 555:246–250.

Rahbek C. 1995. The elevational gradient of species richness: a uniform pattern? Ecography. 18:200–205.

Rambaut A., Drummond A.J., Xie D., Baele G., Suchard M.A. 2018. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Syst. Biol. 67:901–904.

Raven P.H., Gereau R.E., Phillipson P.B., Chatelain C., Jenkins C.N., Ulloa Ulloa C. 2020. The distribution of biodiversity richness in the tropics. Sci Adv. 6:eabc6228.

R Core Team. 2017. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Reid N.M., Hird S.M., Brown J.M., Pelletier T.A., McVay J.D., Satler J.D., Carstens B.C. 2014. Poor fit to the multispecies coalescent is widely detectable in empirical data. Syst. Biol. 63:322–333.

Ren R., Wang H., Guo C., Zhang N., Zeng L., Chen Y., Ma H., Qi J. 2018. Widespread whole

genome duplications contribute to genome complexity and species diversity in angiosperms. Mol. Plant. 11:414–428.

Revell L.J. 2009. Size-correction and principal components for interspecific comparative studies. Evolution. 63:3258–3268.

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

Rice A., Glick L., Abadi S., Einhorn M., Kopelman N.M., Salman-Minkov A., Mayzel J., Chay O., Mayrose I. 2015. The Chromosome Counts Database (CCDB) - a community resource of plant chromosome numbers. New Phytol. 206:19–26.

Ricklefs R.E., Latham R.E., Qian H. 1999. Global patterns of tree species richness in moist forests: distinguishing ecological influences and historical contingency. Oikos. 86:369–373.

Rose J.P., Kleist T.J., Löfstrand S.D., Drew B.T., Schönenberger J., Sytsma K.J. 2018. Phylogeny, historical biogeography, and diversification of angiosperm order Ericales suggest ancient Neotropical and East Asian connections. Mol. Phylogenet. Evol. 122:59–79.

Salazar L., Homeier J., Kessler M., Abrahamczyk S., Lehnert M., Krömer T., Kluge J. 2015. Diversity patterns of ferns along elevational gradients in Andean tropical forests. Plant Ecol. Divers. 8:13–24.

Salman-Minkov A., Sabath N., Mayrose I. 2016. Whole-genome duplication as a key factor in crop domestication. Nat Plants. 2:16115.

Sang W. 2009. Plant diversity patterns and their relationships with soil and climatic factors along

an altitudinal gradient in the middle Tianshan Mountain area, Xinjiang, China. Ecol. Res. 24:303–314.

Santamaría-Aguilar D., Monro A.K. 2019. Compendium of *Freziera* (Pentaphylacaceae) of South America including eleven new species and the typification of 22 names. Kew Bull. 74:14.

Särkinen T., Staats M., Richardson J.E., Cowan R.S., Bakker F.T. 2012. How to open the treasure chest? Optimising DNA extraction from herbarium specimens. PLoS One. 7:e43808.

Sayyari E., Mirarab S. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. Mol. Biol. Evol. 33:1654– 1668.

Shen X.-X., Salichos L., Rokas A. 2016. A genome-scale investigation of how sequence, function, and tree-based gene properties influence phylogenetic inference. Genome Biol. Evol. 8:2565–2580.

Simmons M.P., Sloan D.B., Gatesy J. 2016. The effects of subsampling gene trees on coalescent methods applied to ancient divergences. Mol. Phylogenet. Evol. 97:76–89.

Smith M.L., Hahn M.W. 2021a. New approaches for inferring phylogenies in the presence of paralogs. Trends Genet. 37:174–187.

Smith M.L., Hahn M.W. 2021b. The frequency and topology of pseudoorthologs. bioRxiv.:2021.02.17.431499.

Smith S.A., Brown J.W., Walker J.F. 2018. So many genes, so little time: A practical approach

to divergence-time estimation in the genomic era. PLoS One. 13:e0197433.

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

Solís-Lemus C., Bastide P., Ané C. 2017. PhyloNetworks: a package for phylogenetic networks. Mol. Biol. Evol. 34:3292–3298.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30:1312–1313.

Štorchová H., Hrdličková R., Chrtek J. Jr, Tetera M., Fitze D., Fehrer J. 2000. An improved method of DNA isolation from plants collected in the field and conserved in saturated NaCl/CTAB solution. Taxon. 49:79–84.

Streicher J.W., Schulte J.A. 2nd, Wiens J.J. 2016. How should genes and taxa be sampled for phylogenomic analyses with missing data? An empirical study in iguanian lizards. Syst. Biol. 65:128–145.

Sun L., Meng K., Liao B., Li C., Zhang Y., Liao W., Chen S. 2017. Development and Characterization of Genomic SSR Markers for *Anneslea fragrans* (Pentaphylacaceae). Appl. Plant Sci. 5:1700086.

Tsou C.-H., Li L., Vijayan K. 2016. The intra-familial relationships of Pentaphylacaceae s.l. as revealed by DNA sequence analysis. Biochem. Genet. 54:270–282.

Ulloa C.U., Zarucchi J.L., León B. 2004. Diez años de adiciones a la flora del Perú: 1993-2003.

Arnaldoa. Edición Especial:1–242.

Vargas O.M., Heuertz M., Smith S.A., Dick C.W. 2019. Target sequence capture in the Brazil nut family (Lecythidaceae): Marker selection and in silico capture from genome skimming data. Mol. Phylogenet. Evol. 135:98–104.

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

Weitzman A.L. 1987. Taxonomic studies in *Freziera* (Theaceae), with notes on reproductive biology. J. Arnold Arbor. 68:323–334.

Weitzman A.L. 1988. Systematics of *Freziera* Willd. (Theaceae). .

Weitzman A.L., Dressler S., Stevens P.F. 2004. Ternstroemiaceae. In: Kubitzki K., editor. Flowering Plants. Dicotyledons: Celastrales, Oxalidales, Rosales, Cornales, Ericales. Berlin, Heidelberg: Springer Berlin Heidelberg. p. 450–460.

Wickett N.J., Mirarab S., Nguyen N., Warnow T., Carpenter E., Matasci N., Ayyampalayam S., Barker M.S., Burleigh J.G., Gitzendanner M.A., Ruhfel B.R., Wafula E., Der J.P., Graham S.W., Mathews S., Melkonian M., Soltis D.E., Soltis P.S., Miles N.W., Rothfels C.J., Pokorny L., Shaw A.J., DeGironimo L., Stevenson D.W., Surek B., Villarreal J.C., Roure B., Philippe H., dePamphilis C.W., Chen T., Deyholos M.K., Baucom R.S., Kutchan T.M., Augustin M.M., Wang J., Zhang Y., Tian Z., Yan Z., Wu X., Sun X., Wong G.K.-S., Leebens-Mack J. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc. Natl. Acad. Sci. U. S. A. 111:E4859–68.

Wolf P.G., Robison T.A., Johnson M.G., Sundue M.A., Testo W.L., Rothfels C.J. 2018. Target sequence capture of nuclear-encoded genes for phylogenetic analysis in ferns. Appl. Plant Sci. 6:e01148.

Xu B., Yang Z. 2016. Challenges in species tree estimation under the multispecies coalescent model. Genetics. 204:1353–1368.

Yang Y., Smith S.A. 2014. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. Mol. Biol. Evol. 31:3081–3092.

Yan Z., Smith M.L., Du P., Hahn M.W., Nakhleh L. 2021. Species tree inference on data with paralogs is accurate using methods intended to deal with incomplete lineage sorting. bioRxiv.:498378.

Yu Y., Harris A.J., Blair C., He X. 2015. RASP (Reconstruct Ancestral State in Phylogenies): a tool for historical biogeography. Mol. Phylogenet. Evol. 87:46–49.

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

Zhang C., Scornavacca C., Molloy E.K., Mirarab S. 2020. ASTRAL-Pro: quartet-based species-tree inference despite paralogy. Mol. Biol. Evol. 37:3292–3307.

Zizka A., Steege H.T., Pessoa M. do C.R., Antonelli A. 2018. Finding needles in the haystack: where to look for rare species in the American tropics. Ecography. 41:321–330.

FIGURE CAPTIONS

FIGURE 1. Morphological diversity and geographic distribution of *Freziera*. Photos at left illustrate leaf diversity in *Freziera*, their habit as shrubs and trees, and typical flower and fruit morphology: a) *F. guatemalensis*, b) *F. cyanocantha*, c) *F. candicans*, d) *F. dudleyi*, e) *F. microphylla*, f) *F. lanata*, g) *F. grandiflora*, h) *F. minima*. Map at left (i) illustrates its distribution, with highest density in montane regions of the Neotropics. Photos: a,c,h) Daniel Santamaría-Aguilar; b,d) Robin Foster; e) Alwyn H. Gentry; f) Chris Davidson; g) Alvaro J. Pérez Castañeda.

FIGURE 2. Example gene trees illustrate a) a gene tree with high bipartition support and a topology consistent with a single-copy gene and b) a gene tree with low bipartition support and a topology consistent with a cryptic paralog. Text color of tip names reflects clade assignments outlined in Figure 3. Cryptic paralogs, which result from a combination of biological factors including gene and genome duplication and artifacts of herbiomic data quality, were common in the hybrid-enriched target capture phylogenomic dataset of *Freziera*.

FIGURE 3. Phylogenetic relationships within *Freziera*. Species tree topologies and support values for a) the ASTRAL-Pro all.orthologs+para analysis trimmed to one individual per species and b) the *BEAST analysis. Support values represent local posterior probabilities (ppl) and posterior probabilities for a) and b), respectively. Nodes that were constrained as monophyletic in the *BEAST analysis are indicated with an asterisk (❊). Colors throughout correspond to nine newly named clades. Tips are connected with dashed lines to indicate areas of conflict between species tree analyses. Cartoon topologies in (c) summarize major topological differences that were

common across different datasets as described beneath each tree. Note that each cartoon topology represents one possible outcome, but not necessarily all disagreements recovered (e.g. the Arbutifolia clade was twice found sister to *F. grisebachii* instead of forming a grade as pictured on the far right). Additionally, species trees may contain multiple of these conflicts simultaneously; Table 2 summarizes the degree of topological conflict with the consensus for each analysis.

FIGURE 4. Biogeographic reconstruction using the DEC+J model along the *BEAST species tree resolves the northern Andes as an ancestral and source region for *Freziera*. The map at left shows the areas defined (Mesoamerica= yellow; northern Andes = orange; central Andes = magenta; Guiana Shield = light purple). Distribution of each species in the defined areas is presented in boxes at right of the phylogeny. The bubble graph shows the frequency and direction of movement between areas, while the graph at bottom shows a lineage through time plot (black line) for *Freziera* as a schematic of Andean elevation through time in the northern (orange) and central (pink) Andes.

FIGURE 5. Map of closely-related central Andean species of *Freziera* distributed in climatically similar, but geographically disjunct areas. Dashed lines in shades of the color used to highlight their respective clade (see Fig. 3) connect species at tips of the *BEAST phylogeny to their respective occurrence points and highlight this repeated pattern of geographic separation across the phylogeny.

FIGURE 6. Scatterplots illustrating a greater occupancy of environmental niche space by Northern Andean species of *Freziera* relative to species from other biogeographical regions. Relationships between (a) climate PC1 versus PC2 and (b) soil PC1 versus PC2 from a phylogenetic principal components analysis (pPCA), and (c) average latitude versus elevation are shown. Text in corners of pPCA plots describes the separation of variables. Color schemes reflect biogeography and correspond to regions outlined in Figure 4.

TABLE 1. Names applied to each of the 22 datasets (11 without and 11 with paralogs) and used throughout the text, descriptions of the criteria used and thresholds set for gene or gene tree filtering, and the number of orthologous loci selected by each, with the total number of loci after the addition of 31 known paralogs in parentheses.

TABLE 2. Summary of results for the 22 ASTRAL analyses of datasets following different filtering criteria, without and with the inclusion of paralogs. Dataset names follow those defined in Table 1. Results from analyses without paralogs are listed on the left side of each column, results from those including paralogs are present on the right in parentheses. Metrics summarize gene tree concordance (Normalized Quartet Score), support (average localized posterior probability (ppl) at nodes and the proportion of nodes with ppl≥0.95), and RF distance of the species relative to the consensus topology. Values for the three best-performing datasets are bolded for each metric. The four columns on the right summarize major topological conflicts (illustrated in Fig. 3) between the species tree and the consensus topology: bolded plus signs (+) indicate that the species tree agrees with the consensus topology, minus signs (-) indicate disagreement, and “n/a” is reported for branching order of Elaphoglossifolia group in instances

where the Elaphoglossifolia group was not resolved as monophyletic. Asterisks for disagreement between *clocklike.bipartition* and *low.%.internal+para* and the consensus resolution of two clades in the Candicans group mark disagreement by the alternative placement of only one species.

SUPPLEMENTARY TABLE 1. Per sample data--species name, voucher information and herbarium code for tissue gathered from specimens (codes follow Index Herbariorum: http://sweetgum.nybg.org/science/ih/), provenance, and sample ID used in phylogenetic analyses--and summary statistics generated with HybPiper. Detailed descriptions of these columns are available at (https://github.com/mossmatters/HybPiper/wiki).

SUPPLEMENTARY TABLE 2. Per locus statistics for alignments including outgroup sequences (columns A-V) and excluding outgroup sequences (columns W-AN). Root-to-tip variation, tree length, bipartition support, average bootstrap values, and percent of internal branch lengths in the total tree length (columns B-F) were calculated from gene trees.

SUPPLEMENTARY TABLE 3. Per species values for environmental variables, including principal component (PC) scores for the first two climate and soil PCs; averages for 19 climatic variables, elevation and 12 soil variables; and minimum, maximum, and average latitude.

SUPPLEMENTARY FIGURE 1. Topologies for a) the consensus tree and ASTRAL species trees for each of the 22 datasets: b) *clocklike.bipartition*, c) *tree.length*, d) *PI.per.branch*, e) *high.%.internal*, f) *proportion.PI*, g) *average.BS*, h) *low.%.internal*, i) *bipartition*, j)

variable.sites, k) *1000bp*, l) *all.orthologs*, m) *clocklike.bipartition+para*, n) *tree.length+para*, o) *PI.per.branch+para*, p) *high.%.internal+para*, q) *proportion.PI+para*, r) *average.BS+para*, s) *low.%.internal+para*, t) *bipartition+para*, u) *variable.sites+para*, v) *1000bp+para*, w) *all.orthologs+para*. Node values represent local posterior probabilities (ppl); colors of tip labels reflect clade assignments (see Fig 3).

SUPPLEMENTARY FIGURE 2. Phyparts summaries showing high discordance between gene trees and the a) *all.orthologs*, b) *bipartion*, and c) *clocklike.bipartition* species trees; datasets include 314, 166, and 131 loci, respectively. Numbers on branches indicate the number of genes concordant with the species tree at that node (top), and the number in conflict with that clade (bottom). Pie charts show the proportion of genes that support the species tree topology (blue), the proportion that support the main alternative for that clade (green), the proportion that support the remaining alternatives (red), and the proportion that have less than 50 % bootstrap support (grey).

SUPPLEMENTARY FIGURE 3. Time-calibrated phylogeny from *BEAST. Node values represent node ages in millions of years (Ma); blue bars at nodes represent the 95% highest probability density (HPD) of node ages.