Strong phylogenetic signal despite high phylogenomic complexity in an Andean plant radiation (*Freziera,* Pentaphylacaceae)

Laura Frost[1,2], Ana M. Bedoya[1], & Laura Lagomarsino[1]

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

[2]Biology Department, University of South Alabama, Mobile, AL 36688

ABSTRACT

The Andes mountains of western South America are a globally important biodiversity hotspot, yet there is a paucity of resolved phylogenies for plant clades from this region. Filling an important gap to our understanding of the World's richest flora, we present the first phylogeny of *Freziera* (Pentaphylacaceae), an Andean-centered, cloud forest radiation. Our dataset was obtained via hybrid-enriched target sequence capture of Angiosperms353 universal loci for 50 of the ca. 75 spp., obtained almost entirely from herbarium specimens. We identify high phylogenomic complexity in *Freziera*, including a significant proportion of paralogous loci and a high degree of gene tree discordance. Via gene tree filtering, by-eye observation of gene trees, and detailed examination of warnings from recently improved assembly pipelines, we identified that cryptic paralogs (i.e., the presence of only one copy of a multi-copy gene due to assembly errors) were a major source of gene tree heterogeneity that had a negative impact on phylogenetic inference and support. These cryptic paralogs likely result from limitations in data collection that are common in museomics, combined with a history of genome duplication; they may be common in plant phylogenomic datasets. After accounting for cryptic paralogs as source of gene tree error, we identified a significant, but non-specific signal of introgression using Patterson's D and f4 statistics. Despite phylogenomic complexity, we were able to resolve *Freziera* into nine well-supported subclades whose histories have been shaped by myriad evolutionary processes, including incomplete lineage sorting, historical gene flow, and gene duplication. Our results highlight the complexities of plant phylogenomics, and point to the need to test for multiple sources of gene tree discordance via careful examination of empirical datasets.

KEYWORDS: Andean radiation; Angiosperms353; gene tree discordance; gene tree filtering; introgression; locus filtering; museomics; paralogy

The Andean mountains in South America are one of the most species-rich areas of the world and serve as a center of diversity for many plant groups (Gentry 1982; Mutke and Barthlott 2005). The recent uplift of the Andes has resulted in some of the fastest plant evolutionary radiations reported to date (Madriñán et al. 2013; Hughes 2016), with some greatly phenotypically diverse clades (Hughes and Eastwood 2006; Lagomarsino et al. 2016). With a significant portion of Andean plant biodiversity still yet to be described (Ulloa Ulloa et al. 2017), many— if not most— Andean plant species have never been included in a phylogeny. However, these phylogenies are fundamental toward understanding the evolutionary patterns that contribute to the origin and diversification of biodiversity in the World's richest flora.

Establishing well-supported phylogenies for Andean plant groups is challenging for many reasons. Short divergence times between speciation events, incomplete lineage sorting (ILS), incipient speciation, and introgression all contribute to poor phylogenetic resolution and high discordance between gene trees and species trees (Vargas et al. 2017; Morales-Briones et al. 2018; Lagomarsino et al. 2022). This is further complicated by repeated whole genome duplication events throughout the evolutionary history of plants, at both deep (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. It is difficult to achieve full taxon sampling as species are often distributed in remote locations, and members of clades occur in many countries, each with different policies concerning collection and exportation of samples. As a result, achieving dense sampling of Andean-centered lineages commonly requires the use of herbarium specimens as a source of genetic material, which is associated with lower quality and quantity DNA than in freshly collected or silica dried leaf tissue (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 previously intractable groups. Advancements in genomic sequencing, including hybrid-enriched target sequence capture, allow for the collection of hundreds to thousands of loci, even from degraded DNA from natural history collections in museomic studies (Bakker 2017; McKain et al. 2018). Further, the development of universal probe sets facilitates sequencing of hundreds of loci for any system, regardless of genomic resources available (Johnson et al. 2019). Analytical methods are also increasingly able to accommodate multiple biological sources of gene tree discordance (Ogilvie et al. 2017; Solís-Lemus et al. 2017; Zhang et al. 2018).

Still, the efficient identification of multi-copy loci, especially paralogs, remains a barrier to phylogenomics in many plant clades. Several processes result in multiple sequences for a

single locus within an individual: allelic variation, gene copies created by a duplication event within the same genome (i.e., paralogs), and orthologs that were reunited in the same genome by allopolyploidization (i.e., homeologs). Automated detection of multi-copy genes can be hindered when sequence data fails to meet contig number, contig length, or read depth thresholds set during assembly. The presence of only one copy of a multi-copy gene in a sequenced dataset (i.e., hidden paralogy) may either be due to biological processes or as data artifacts. “Pseudo-orthologs” are a type of hidden paralogy that results from a biological process: differential loss after gene duplication that leads to the presence of a single but non-orthologous copy across species in nature (Smith and Hahn 2021a, 2021b). While there is some evidence that coalescent-based phylogenetic methods are relatively robust to pseudo-orthologs (Smith and Hahn 2021a, 2021b), this finding is applicable only under certain conditions that do not include whole genome duplication followed by rediploidization— a phenomenon common in plants (Li et al. 2021). Hidden paralogy can also be artifactual, including when differential assembly of copies results in sequences made up of non-orthologous contigs (chimeric sequences; https://github.com/mossmatters/HybPiper). We refer to these as “cryptic paralogs”. The presence of cryptic paralogs in phylogenomic datasets and their impact on results is underexplored.

We tackle the many challenges of Andean plant phylogenomics to infer the first phylogeny and introgression history of *Freziera* (Pentaphylacaceae), a group that previously lacked any phylogenetic information. *Freziera* is a genus of 75 spp. of trees and shrubs that is widely distributed throughout montane regions of the Neotropics, from southern Mexico to Bolivia, with a center of diversity in Andean cloud forests (61 spp. are found in this biome;(Santamaría-Aguilar and Monro 2019); Fig. 1). Unlike many charismatic cloud forest radiations that have dramatic variation in floral morphology, *Freziera* species are consistent in life history traits related to reproduction: species are dioecious, flowers exhibit a generalist invertebrate pollination syndrome, and all species produce fleshy berries (Weitzman 1987, 1988; Weitzman et al. 2004; Santamaría-Aguilar and Monro 2019). There 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 leaf size varies over 400-fold within the genus. Ploidy levels in *Freziera* are unknown due to few genomic resources (i.e., only 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)]). There has been at least one whole genome duplication (WGD) event in an ancestor of Pentaphylacaceae at a deep node in the inclusive order Ericales (Larson et al. 2020), and

available chromosome counts suggest either more recent duplication events or loss of copies across tax (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]).

Using Angiosperms353 (Johnson et al., 2019) target enrichment data derived almost entirely from herbarium specimens, we establish a phylogenetic baseline of *Freziera* and identify historical introgression in the group. We further explore how locus filtering impacts species tree resolution and support, as well as gene tree discordance. We identify widespread cryptic paralogs, and provide suggestions to explore complex empirical phylogenomics datasets, especially those with a past history of genome duplication and that are obtained with a high reliance on natural history collections.

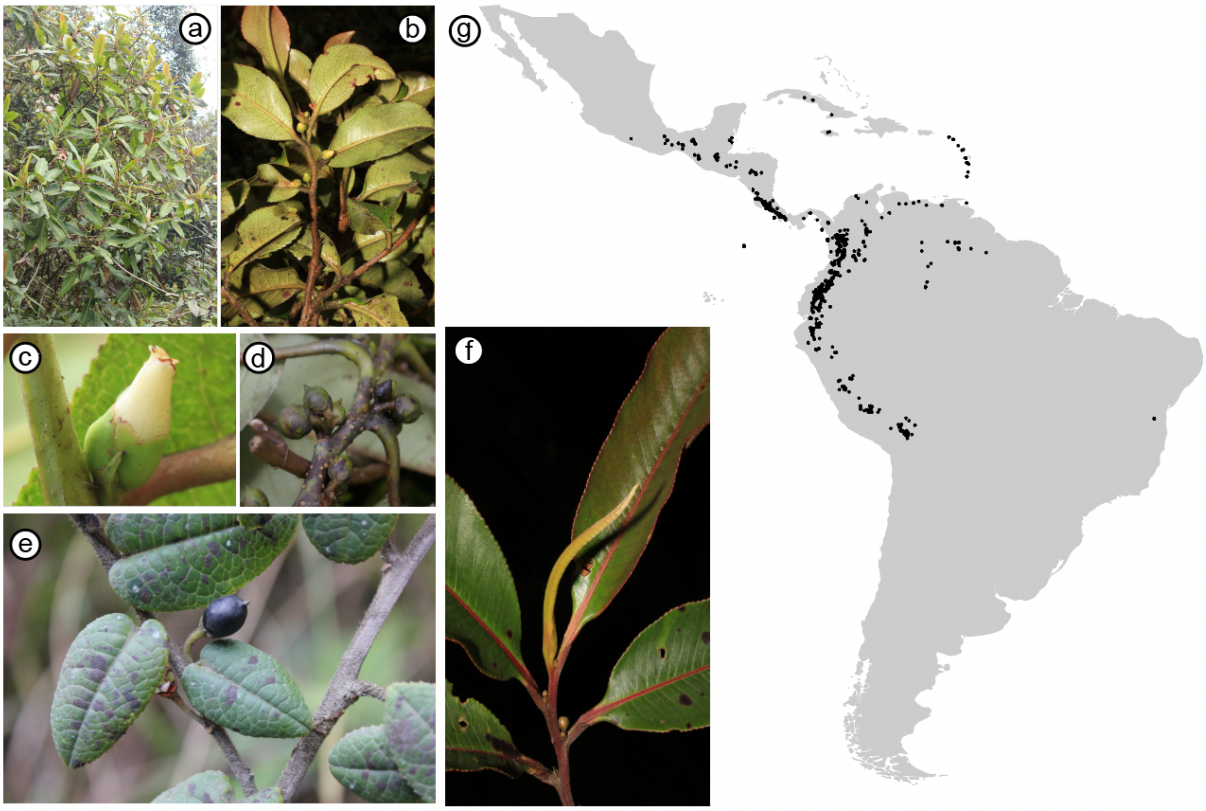

Figure 1. Morphological and geographic diversity of *Freziera*. Photos at left illustrate leaf diversity in Freziera, their habit as shrubs (a), leaf variation (b-f), typical flower (b, c) and fruit (d, e) morphology, and the characterisitic apical bud (f). Map at right illustrates *Freziera*'s distribution, with highest density in montane regions of the Neotropics. Photos by L. Lagomarsino.

MATERIALS AND METHODS

Taxon Sampling

119 Ninety-three accessions representing 55 *Freziera* species—approximately 73% of the species
120 diversity of the genus—were sampled for the ingroup. All but five *Freziera* samples came from
121 herbarium specimens, which had an average age of 31.6 (range: 7.4– 82.8 years; Table S1).
122 Nine accessions from other genera in Pentaphylacaceae were sampled for the outgroup,
123 including *Eurya japonica*, from the genus sister to *Freziera*; *Cleyera albopunctata,* a member of
124 tribe Frezierieae; and 7 species of *Ternstroemia,* belonging to the sister tribe Ternstroemieae
125 (Weitzman et al. 2004; Tsou et al. 2016).
126
127 *DNA extraction, library prep, target enrichment, and sequencing*
128 Detailed descriptions of laboratory methods are provided in Appendix 1. Briefly, dried leaf
129 tissue, primarily from herbarium specimens, was homogenized using a FastPrep-24TM 5G (MP
130 Biomedicals, Solon, Ohio, United States). DNA extraction followed a modified sorbitol extraction
131 protocol (Štorchová et al. 2000). Library preparation was carried out using KAPA Hyper Prep
132 and KAPA HiFi HS Library Amplification kits (F. Hoffmann-La Roche AG, Basel, Switzerland)
133 with iTru i5 and i7 dual-indexing primers (BadDNA, University of Georgia, Athens, Georgia,
134 United States). Target enrichment was carried out using the MyBaits Angiosperms353 universal
135 probe set (Däicel Arbor Biosciences, Ann Arbor, MI;(Johnson et al. 2019; Hale et al. 2020)).
136 DNA libraries were sent to Novogene Corporation Inc., (Sacramento, California, United States)
137 for sequencing in one lane of an Illumina Hiseq 3000 platform (San Diego, California, United
138 States) with 150 bp paired-end reads.
139
140 *Raw data processing and locus extraction and alignment cleaning*
141 Demultiplexed raw sequence reads were trimmed with illumiprocessor v2.0.9 (Faircloth
142 2013, 2016), a wrapper for trimmomatic v0.39 (Bolger et al. 2014). Default settings were used
143 and reads with a minimum length of 40 bp kept. Trimmed reads were assigned to their target
144 genes and initially assembled with HybPiper v1.3.1 (Johnson et al. 2016); see Appendix 1 for
145 detailed description (Johnson et al. 2016). Examination of resulting alignments and gene trees
146 led to suspicion of artifactual errors in the dataset, mainly cryptic paralogs (see below; Fig. S1).
147 To test the effect of relaxing thresholds on paralog detection, a subset of samples and loci
148 were selected for rapid assembly. This subset included the six accessions of *F. candicans*,
149 represented by two silica-dried and four herbarium samples of varying quality, and the first 15
150 numbered loci in the list of the the Angiosperm353 targeted genes, which included one flagged
151 paralog and seven suspected cryptic paralogs in the initial run (Table S3). The coverage cutoff
152 for Spades was relaxed from the default of 8 to 6, and the minimum length percentage for a

153 contig to be considered as a paralog was relaxed from 75% of the reference protein length to
154 50% (Table S4). The recent release of HybPiper v.2.0.1 promises improved paralog detection
155 and greater transparency of the assembly process via reporting of stitched contigs
156 (https://github.com/mossmatters/HybPiper/). Therefore, tests on the 15-locus *F. candicans*
157 subset were run with both HybPiper v.1.3.1 and HybPiper v.2.0.1. We summarized the output of
158 HybPiper v.2.0.1 results by compiling (1) the proportion of sequences with stitched contigs per
159 locus and (2) which genes triggered a paralog warning due to contig depth for each sample
160 (Table S3).
161
162 *Phylogenetic analyses*
163 *Gene tree inference and filtering*.—Details of data processing prior to final gene tree
164 inference are available in Appendix 1. The following methods were performed on supercontig
165 alignments and gene trees resulting from HybPiper v.1.3.1 assemblies with default settings,
166 except where stated. Summary statistics for alignments (with outgroups removed) were
167 obtained using AMAS (Borowiec 2016), including alignment length, number of variable sites,
168 and proportion of parsimony informative sites. Genes with fewer than 25 ingroup samples were
169 excluded from further analyses. Gene trees were inferred with IQ-TREE multicore
170 v2.1.1(Nguyen et al. 2015) combining model section via ModelFinder (Kalyaanamoorthy et al.
171 2017), tree search, 5000 ultrafast bootstrap replicates (Hoang et al. 2018), and 5000
172 Shimodaira-Hasegawa-like approximate likelihood-ratio tests (SH-aLRT;(Guindon et al. 2010;
173 Anisimova et al. 2011)). Average bootstrap support and percent of tree length made up by
174 internal branches were extracted from the IQtree output for each gene tree.
175 Observation of gene trees and alignments suggested the presence of more non-
176 orthologous loci in the dataset than were flagged by the paralog_retriever.py script. Evidence for
177 this included split clades of outgroup species and distinctive motifs in alignments that reflected
178 higher sequence divergence than expected by allelic variation alone. With the assumption that
179 coalescent methods are robust enough to noise introduced by cryptic paralogs (Markin and
180 Eulenstein 2020; Legried et al. 2021; Smith and Hahn 2021b; Yan et al. 2021), we generated a
181 preliminary species tree with ASTRAL-III (Zhang et al. 2018) from the 314 maximum likelihood
182 (ML) gene tree topologies. Gene trees were then examined by eye in FigTree v.1.4.3 (Rambaut
183 2014) and sorted as hypothesized single copy orthologs or cryptic paralogs. Gene trees in
184 which ingroup species group into two subclades containing samples of the same species and
185 separated by relatively long internal branch lengths were considered cryptic paralogs in this by-
186 eye identification method (Fig. S1).

To explore how locus filtering impacts gene tree-species tree discordance and species tree resolution and support, genes and gene trees were then filtered using eleven empirical criteria to exclude potential sources of gene tree estimation error (Table 1). Filtering criteria included four alignment-based metrics: (1) alignment length (cutoff: >1000 bp; Liu et al. 2015), (2) number of variable sites (>750), (3) proportion of parsimony informative sites (0.075), and (4) the number of parsimony informative sites relative to the number of unrooted internal branches (# of parsimony informative sites/(# of tips-3); cutoff: 4x number of internal branches). Except for alignment length, cutoff thresholds near the average value in our dataset were selected to maintain a similar number of loci for each criterion (i.e., ~150 loci). Three filtering criteria applied to tree length: (1) above average overall tree length and both (2) above and (3) below average percent of tree length made up by internal branches. Sets of gene trees with both above average and below average percentages of internal branch lengths were selected because, while a high percentage of internal branch length can signal high phylogenetic signal (Shen et al., 2016), it can also indicate pseudo-orthologs (Smith and Hahn, 2016b)--and, likely, cryptic paralogs. Finally, three filters related to gene tree support were applied: (1) above average bootstrap support across all nodes, (2) above average bipartition support compared to a species tree, and (3) below average root-to-tip variance (i.e., clock-likeness). The 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. Bipartition support was calculated against the ASTRAL-III tree inferred from the best maximum likelihood (ML) topologies of 314 gene trees. For below average root-to-tip variance, loci that had lower than average root-to-tip variance, or greater clock-likeness, from the set with above average bipartition support genes were selected; this filtering scheme was based on recommendations in Smith et al. (2018).

Finally, to explore how automated paralog detection compares to other methods of gene tree filtering, gene trees generated in HybPiper v.1.3.1 were filtered by whether they were flagged as potentially paralogous by HybPiper v.2.0.1 in two ways. First, loci that had a long paralog warning (i.e., multiple long contigs for one or more samples, equivalent to the metric used to flag paralogs in HybPiper v.1.3.1) were excluded. Second, loci that received any kind of paralog warning, either for long-contigs or for contig depth, were excluded. These filters were also applied to the HybPiper v.2.0.1 gene trees and compared to the HybPiper v.1.3.1 datasets, though this is a less direct comparison due to differences in assembly.

Table 1. Names applied to each of the datasets used throughout the text, descriptions of the criteria used and thresholds set for gene or gene tree filtering, and the number of loci selected by each.

Dataset name	Description	Number of loci
clocklike.bipartition	gene trees with above average bipartition support and below average root-to-tip variance	131
tree.length	total tree length above average	140
PI.per.branch	number of parsimony informative sites at least 4x the number of internal branches	148
high.%.internal	proportion of total tree length comprising internal branch lengths above average	149
proportion.PI	alignment contains at least 7.5% parsimony informative sites	160
average.BS	average bootstrap support for gene tree above average	163
low.%.internal	proportion of total tree length comprising internal branch lengths below average	165
bipartition	gene trees with above average bipartition support relative to a preliminary species tree	166
variable.sites	alignment contains at least 750 variable sites in	168
by.eye	genes determined to be single copy after by-eye inspection of gene trees and alignments	182
hybpiper2.no.warnings	loci without any kind of paralog warning from HybPiper v.2.0.1	187
hybpiper2.orthologs	loci without long paralog warnings by HybPiper v.2.0.1	261
1000bp	alignment at least 1000 bp	300
all.orthologs	all loci not flagged as paralogs by HybPiper v.1.3.1	314

Species tree inference and exploration of gene tree discordance.—Accessions present in <10% of processed gene trees were trimmed using the R package ape (Paradis and Schliep 2019). Maximum likelihood trees generated by IQtree (see Methods above) were used to estimate species trees in ASTRAL-III (Zhang et al. 2018). Sixteen datasets were analyzed: one including all loci not flagged as paralogous by HybPiper v.1.3.1, two including loci without long paralog warnings from HybPiper v.2.0.1 (the same filter was applied to HybPiper v.1 and HybPiper v.2 gene trees), two including loci without any paralog warnings from HybPiper v.2.0.1 (same as previous), and each of the eleven filtered datasets (Table 1).

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*, *by.eye*, and *HybPiper.2.no.warnings* 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 Robinson-Foulds (RF) distances were calculated as summary statistics for species tree inference. A majority-rule consensus tree was generated from the species trees for the 14 HybPiper v.1.3.1 datasets 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).

241

242 *Signals of introgression in Freziera.*—To assess genomic evidence of introgression in
243 species of *Freziera*, we calculated Patterson's D and f4 statistics using the Dtrios function in
244 Dsuite v.0.4 r43 (Malinsky et al. 2021). The VCF file used as input for Dsuite was generated
245 with dDocent (Puritz et al. 2014) by mapping all *Freziera* trimmed reads to the sequences of *T.*
246 *tepezapote* (reference) for the loci in the *bipartition*, *by.eye*, and *hybiper2.no.warnings* datasets.
247 We called SNPs using default values for all mapping parameters. The resulting VCF files were
248 filtered with vcftools (Danecek et al. 2011). We retained SNPs with <50% missing data (--max-
249 missing 0.5) and retained only one SNP per 100 bp window to decrease the likelihood of
250 including linked SNPs (--thin 100). The statistical significance of the D and f4 was assessed
251 using block jackknife on windows of 75–78 SNPS followed by Benjamini-Hochberg correction as
252 implemented in R (Benjamini and Hochberg 1995) to assess family-wise error rate (following
253 (Malinsky et al. 2018). The D and f4 statistics were estimated for all possible trios across the
254 three datasets (13,244, 15,180, 14,190 trios for the *bipartition*, *by.eye*, and
255 *hybiper2.no.warnings* respectively). The ASTRAL species trees inferred from each of the three
256 datasets were specified in Dsuite so that the D and f4 estimated values were arranged
257 according to the tree.
258

259 **RESULTS**

260 *Taxon sampling and target sequence capture success*

261 Voucher information for accessions and per sample data for read trimming and contig
262 assembly are available in Supplementary Table 1. On average, 440 sequences were recovered
263 for 241 loci per sample (range: 0-348). All 353 loci had sequences for at least four samples,
264 and at least 25% of samples were present in 346 loci; per locus summary statistics are available
265 in Table S2. Seventy-six accessions representing 50 of the 75 species of *Freziera* were
266 included in the final ASTRAL-III analyses after the rounds of processing described (see
267 methods; Supplementary Table 1).
268

269 *Phylogenetic analyses*

270 *Chimeric sequence and paralog detection.*—No genes were flagged by HybPiper v.2.0.1
271 as being putative chimeric sequences. Yet, of all 353 loci, only 13 did not include any
272 sequences with stitched contigs. These 13 included six that were removed from further analysis:
273 two with flagged paralogs and four with fewer than 25 ingroup samples. On average, 61.3% of

supercontigs per locus were composed of stitched contigs (range lowest to highest with stitched contigs: 1.2-97.5%).

HybPiper v.1.3.1 flagged 31 loci as paralogous. Of the 322 loci that were not flagged as paralogs by HybPiper v.1.3.1, eight were removed because they contained fewer than 25 ingroup samples, leaving 314 putative “orthologs”. Changes to thresholds on a subset of data had little impact on paralog detection with HybPiper v.1.3.1; the run with relaxed coverage cutoff and paralog minimum length percentage thresholds flagged 32 loci as paralogous. In tests with HybPiper v2.0.1, however, lowering the coverage cutoff to 6 increased the number of genes and/or the length of sequences recovered for some samples (Supplement ##). Lowering the paralog minimum length percentage to 50% increased the number of loci for which paralog warnings were issued. Long paralog warnings were issued for 84 loci by HybPiper v.2.0.1; 79 of these also received paralog warnings by contig depth. An additional 75 loci were not flagged as long paralogs but did receive paralog warning by contig depth for at least one sample, totaling 154 loci with warnings by contig depth and 159 loci with any kind of paralog warning. By eye determination of cryptic paralogs largely corresponded to paralog warnings from HybPiper v.2.0.1 (Table S3).

Gene tree inference and filtering.—Summary statistics for alignments, both including and excluding outgroups, are available in Table S2; gene trees are available on Dryad (*web address added upon acceptance*). 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. Dataset names, thresholds applied, and number of loci meeting thresholds are summarized in Table 1.

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, the *bipartition* dataset resulted in the best overall tree, defined using all criteria except for average node support, where it had the second highest score following *clocklike.bipartition*; *by.eye* and *hybpiper2.no.warnings* species trees also scored high across support criteria (Table 2). Comparisons of loci in these datasets revealed a high degree of overlap in loci (Fig. 2a). Assessment of gene tree concordance with PhyParts revealed high levels of conflict between gene trees and the four species tree topologies explored. Most of the

307 conflict was the result of many different alternative topologies rather than one frequent
308 alternative topology (Fig. 2b, S3).
309 The consensus topology included nine clades that were frequently inferred across
310 analyses: the Humiriifolia clade, *F. grisebachii*, *F. magnibracteolata*, the Canescens clade, the
311 Incana clade, the Arbutifolia clade, the Lanata clade, the Karsteniana clade, and the Calophylla
312 clade (Fig. 3). Six of the sixteen datasets produced species trees consistent with monophyly of
313 the nine clades and their branching order in the consensus tree; an additional three were
314 consistent except for one species (Table 2; Fig. S2a,o,p). Among species trees consistent with
315 the consensus topology, some deep nodes were inferred with high support (ppl≥0.95) across all
316 analyses: the common ancestor of all Freziera, the common ancestor of core Freziera (i.e.,
317 excluding the Humiriifolia clade), the successive node within core Freziera, the Humiriifolia,
318 Canescens, and Incana clades, and the Candicans group (Fig. 3). Species-level relationships

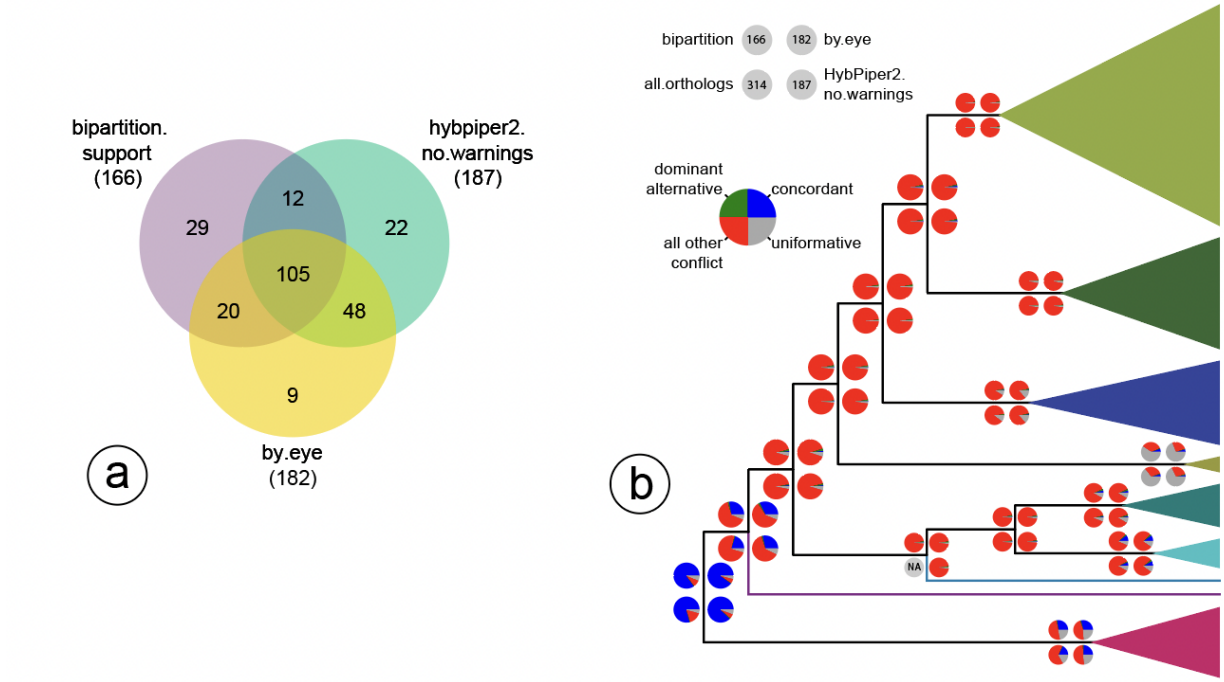

Figure 2. Impact of gene filtering of criteria on gene tree discordance. (a) Venn diagram showing overlap in genes selected by the three top-performing filtering metrics: bipartition support, by eye, and loci without any paralog warnings from HybPiper v.2.0.1. The total number of genes included in each dataset is given in parentheses below the respective dataset name. (b) Phyparts measurements of discordance between gene trees and major nodes of the species tree for (clockwise from the top left) *bipartition*, *by.eye*, *hybpiper2.no.warnings*, and *all.orthologs* datasets. Clade position and color correspond to the consensus topology in Fig. 3.

319 within most subclades were consistent across analyses with the exception of the species-rich
320 Candicans group, in which species-level relationships exhibited differences across analyses.

321 Regions of low support also tended to be in conflict between species trees and the
322 consensus topology. These regions often involved two notable sets of taxa that tend to group
323 together in analyses: the Elaphoglossifolia group, comprising the Incana and Canescens clades
324 and *F. magnibracteolata*; and the Candicans group, which comprises the Calophylla and
325 Karsteniana clades. While the Elaphoglossifolia group was monophyletic in all but five of our
326 analyses (Table 2), though generally not with high support (Fig S2), the unstable placement of
327 *F. magnibracteolata* often rendered it non-monophyletic (Table 2). Contrasting with the
328 Elaphoglossifolia clade, the monophyly of the Candicans group was strongly supported in all
329 analyses, generally with high support; however, the resolution of its constituent Calophylla and
330 Karsteniana subclades was not (Table 2).

Table 2. Summary of results for the 16 ASTRAL analyses of datasets following different filtering criteria. Dataset names follow those defined in Table 1. 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.

Dataset name	Normalized Quartet Score	Average node support	proportion of highly supported nodes	RF distance from consensus	Two clades in Candicans group	Monophyletic Elaphoglossifolia group	Branching order of Elaphoglossifolia group	Placement of Arbutifolia clade
clocklike.bipartition	**0.621**	**0.800**	0.403	30	–*	+	+	+
tree.length	0.524	0.697	0.273	54	–	–	n/a	–
PI.per.branch	0.537	0.709	0.273	50	–	–	n/a	+
high.%.internal	0.541	0.693	0.247	54	–	+	+	+
proportion.PI	0.527	0.718	0.273	48	–	–	n/a	+
average.BS	0.560	0.727	0.312	28	+	+	–	+
low.%.internal	0.586	0.760	0.338	34	+	–	n/a	–
bipartition	**0.601**	**0.792**	**0.455**	**14**	+	+	+	+
variable.sites	0.559	0.746	0.351	28	+	+	+	+
by.eye	**0.634**	**0.792**	**0.442**	24	+	+	+	+
hybpiper2.no.warnings	**0.619**	**0.783**	0.390	22	+	+	+	+
h2.no.warnings.h2gt	0.581	0.771	**0.442**	54	+*	+	+	+
hybpiper2.orthologs	0.582	0.780	0.403	**16**	+	+	+	+
h2.orthologs.h2gt	0.549	0.764	**0.455**	38	+*	+	+	+
1000bp	0.566	0.773	0.429	**16**	+	+	+	+
all.orthologs	0.566	0.777	0.416	**18**	+	–	n/a	+

331

332 *Signals of introgression in Freziera.*—For the *bipartition*, *by.eye*, and
333 *hybpiper2.no.warnings* datasets, p-values for the estimated D statistics could not be calculated
334 for 75.9%, 85.5%, and 93.6%, respectively, of all evaluated trios because of lack of ABBA

BABBA variants (Fig. S4). Estimates of Patterson's D >0 are significant (corrected p-value ≤ 0.05, Z-score >3) for 104 (3.27%), 162 (7.38%), and 46 (5.08%) trios out of the remaining evaluated trios. Resulting f4-ratio statistics across the three datasets also show evidence for multiple instances of introgression in *Freziera* (Fig. S4).

The number of statistically significant D statistics is probably underestimated in our study given the large proportion (~76%–94%) of trios for which p-values could not be calculated due to a lack of ABBA BABBA variants. However, statistical significance of D and F4 statistics are correlated given to shared ancestry among multiple potential introgression donor species (Malinsky et al., 2018). This correlation would likely reduce the proportion of statistically significant D and f4-ratio statistics across datasets in *Freziera*. The f-branch statistic, which incorporates excess allele sharing among taxa in a phylogenetic tree, allows for a better interpretation of introgression patterns given these high correlations (Malinsky et al., 2021). However, simulation analyses have shown that for a relatively small number of unlinked SNPs (<10,000), the proportion of cases where the strongest inferred f-branch signal corresponds to the correct simulated gene flow recipient and donor branches is <20% (Malinsky et al., 2021), much larger than the unlinked SNPs we were able to extract from our targeted sequence capture dataset.

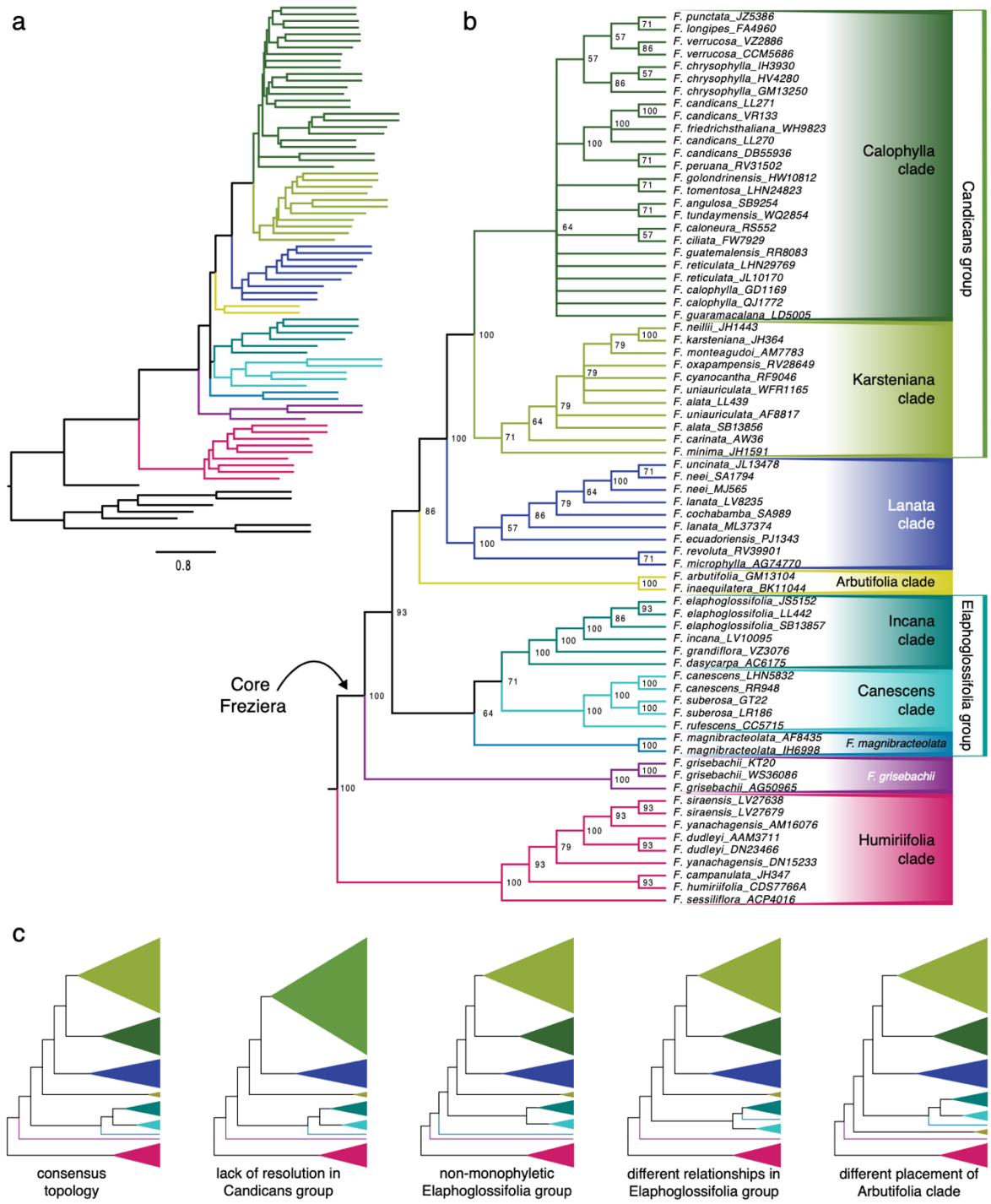

Figure 3. Species tree topology of Freziera, including (a) branch lengths in coalescent units for the bipartition ASTRAL-III tree, (b) the consensus topology for all 14 datasets with bootstrap support for each node and named clades labeled at tips, and (c) cartoon trees of named clades depicting the consensus topology and the most frequent deviations from the consensus (see Table 2 for a comparison of the species tree for each dataset to the consensus tree).

DISCUSSION

Relying on DNA extracted primarily from herbarium specimens, we inferred the first phylogeny of *Freziera* (Pentaphylacaceae), an understudied tropical plant lineage. Our museomic dataset, representing hybrid-enriched target sequence capture of Angiosperms353 loci, highlights the many challenges of working with understudied clades even in the genomic era: no a priori phylogenetic hypothesis, a universal baitset, and poor-quality DNA resulting in short contigs and low coverage in non-coding regions. Despite phylogenomic complexity, we resolve *Freziera* into nine well-supported clades whose histories have been shaped by myriad evolutionary processes, including incomplete lineage sorting, introgression, and gene or genome duplication, likely corresponding to past polyploidy. By further familiarizing ourselves with our data via gene tree filtering, by-eye observation of gene trees, and detailed examination of warnings from recently improved assembly pipelines, we identified that cryptic paralogs were a major source of gene tree heterogeneity that had a negative impact on phylogenetic inference and support. Reducing this spurious signal resulted in a more stable, well-supported species tree. Across analyses, topological differences were primarily concentrated in a few portions of the phylogeny, with backbone relationships largely consistent across analyses, even those including cryptic paralogs.

Freziera phylogenomics

One of the barriers to the study of Neotropical diversification is the difficulty resolving phylogenies of recent, rapid Andean radiations. Despite *Freziera*'s relatively low species richness compared to many cloud forest clades, we find similar hallmarks of explosive radiation in our phylogenetic results. High conflict among gene tree topologies (Fig. S3) was found to underlie a relatively stable species tree for *Freziera* (Fig. 3c). In the face of this complexity, we resolved *Freziera* into nine subclades, most of which are moderately to well supported. Despite their monophyly, these clades generally lack morphological synapomorphies and are not geographically structured. The Humiriifolia clade, which is sister to core *Freziera*, best represents the wide morphological diversity in each subclade: its species have among the largest (*F. humiifolia*) and smallest (*F. yanachagensis*) leaves in the genus, despite occurring in close proximity in the Cordillera del Cóndor region of southern Ecuador. While our phylogenetic backbone is well-supported, there are very high levels of gene tree discordance in our dataset for which we were able to identify three non-mutually exclusive biological sources: introgression, ILS, and gene (or genome) duplication. Despite a genome-wide phylogenomic dataset, we are unable to pinpoint exactly when and where along the phylogeny each of these processes has

occurred. However, our resulting phylogenetic framework provides a robust starting point from which to understand the non-bifurcating nature of diversification of this Andean shrub clade.

Incomplete lineage sorting is common in rapid Andean radiations (Gómez-Gutiérrez et al. 2017; Pouchon et al. 2018; Nicola et al. 2019). Within *Freziera*, the Candicans group, especially its substituent Calophylla clade, carries the hallmarks of ILS due to rapid radiation, including conflicting species relationships with short branch lengths between close relatives (Figs. 3, S2). This results in short coalescent times between speciation events. Not only is gene tree-species tree discordance higher in this clade compared to other regions of *Freziera*'s phylogeny (Fig. 2b, S3), but species relationships also differ significantly across analyses (Fig S2). In fact, most differences in RF distances in Table 2 are due to conflicting relationships in this clade, resulting in multiple polytomies in the consensus tree. Further suggesting a prominent role for ILS in the Calophylla clade, and the Candicans group more broadly, it is the most widespread of *Freziera*'s subclades and includes species with some of the broadest distributions in the genus. These distributional patterns allow a greater possibility that ancestral populations were large and widespread, more likely resulting in ILS (Holder et al. 2001).

As evidenced by multiple significant f4 and D statistics, the history of *Freziera* is also marked by at least some introgression (Fig. S4). Due to limited data from our targeted sequence capture dataset and correlation of the above-mentioned introgression statistics when trios share internal branches, we were not able to isolate the exact branch(es) along which introgression has occurred— a challenge even among the most complete datasets (Tricou et al. 2022). However, we do identify species like *F. magnibracteolata*, whose unstable placement may suggest a potential history of introgression (MacGuigan and Near 2019; Cai et al. 2021). To further assess the extent of gene flow and identify the branches involved in introgression events in this rapid radiation, future research will target a much larger portion of the genome (Malinsky et al., 2021) and will explore non-bifurcating phylogenetic relationships, which likely more accurately reflect phylogenetic relationships in *Freziera*. Unfortunately, computational demands of current phylogenetic network methods are still a barrier to inference of networks on phylogenomic datasets with large number of tips and a relatively small number of independent loci. This challenge is amplified in the absence of a specific hypothesis about which branches were involved in historical introgression.

Finally, either gene or genome duplication has resulted in paralogous loci in *Freziera*. We identify both paralogs for whom multiple copies are identifiable and cryptic paralogs that are represented by only a single copy in our dataset (see discussion below; Fig. S1, Table S3). While the processes that gave rise to paralogs are yet to be examined in detail, it is likely that

423 multicopy paralogous loci are the product of allopolyploidy, especially considering that there is
424 an extensive history of genome duplication via polyploidy in Ericales, the order to which *Freziera*
425 belongs (Larson et al. 2020), as well as chromosome count variation within Pentaphylacaceae
426 (Rice et al. 2015). The signal of introgression we identify may also be due to past polyploidy
427 events.
428
429 *Gene tree filtering & paralog detection*
430 Gene tree estimation error (GTEE) is another source of gene tree heterogeneity that
431 impacts species tree estimation (Molloy and Warnow 2018). Without an explicit method to
432 estimate GTEE in empirical data, multiple criteria, including alignment length, the
433 number/proportion of variable or parsimony informative sites, total tree length, the proportion of
434 internal branch lengths, and average bootstrap support of gene trees, have all been used as
435 implicit proxies to account for GTEE (Leaché et al. 2014; Liu et al. 2015; Shen et al. 2016; Blom
436 et al. 2017). However, when applied to the *Freziera* dataset we find that all criteria but alignment
437 length (i.e., *1000bp* dataset) resulted in species trees that were more unlike the consensus
438 topology relative to the *all.orthologs* dataset and tended to have lower support (Table 2). The
439 better performance of all loci relative to what should be the most informative loci after filtering for
440 low phylogenetic signal suggests that strong, but inaccurate signal is a larger problem than lack
441 of signal in our dataset. Indeed, with an average alignment length over 3,000 bp and an average
442 of almost 5 parsimony informative sites per internal branch for the ingroup in each gene tree
443 (Table S2), lack of phylogenetic information per locus would not seem to be the primary issue in
444 this dataset as it may be for shorter targets, like RADseq data (McKain et al. 2018).
445 While most criteria used to filter gene trees did not result in improvements in
446 phylogenetic performance, filtering gene trees by bipartition support did (Table 2). Specifically,
447 relative to the *all.orthologs* species tree, the *bipartition* species tree had lower gene tree
448 discordance (i.e., higher normalized quartet score) and higher average support (i.e., average
449 node support and proportion of highly supported nodes), and was more similar to the consensus
450 topology (Table 2). While the *all.orthologs* species tree did not recover all major clades identified
451 in the consensus topology (in particular, the Elaphoglossifolia group was not monophyletic), the
452 *bipartition* species tree was consistent with the consensus tree, including monophyly of the nine
453 named clades and their relative branching order. A second filtering step to remove those gene
454 trees with below average root-to-tip variance after first filtering by bipartition support (i.e.,
455 *clocklike.bipartition*) also resulted in a monophyletic Elaphoglossifolia group, but lost full
456 resolution of the two clades within the Candicans group. These results highlight the trade-off

457 between more data versus more-curated data: the reduction of noise through multiple filtering
458 steps improved resolution and support in some areas of the tree (Supplement #,#), but the loss
459 of information resulted in lowered support and resolution in others. Overall, we find that filtering
460 gene trees by the hypothetical accuracy of phylogenetic information via bipartition support
461 resolved relationships among major clades and resulted in higher gene tree concordance and
462 support for species trees—a desirable goal for phylogenetics and macroevolutionary studies,
463 especially those that require a bifurcating phylogeny.

464 Much of the improvement identified in filtering gene trees by bipartition support can be
465 attributed to the removal of cryptic paralogs: many of the loci that we identified as cryptic
466 paralogs in a by-eye examination of our data were removed through this mechanism. These loci
467 further share significant overlap with those that had HybPiper2 paralog warning, and there was
468 significant overlap in the datasets generated by filtering by bipartition support, HybPiper2
469 warnings, and by-eye identification of cryptic paralogs (Fig. 2a). The high degree of overlap
470 between these datasets suggests that (1) filtering by bipartition support effectively filters out loci
471 that receive paralog warnings and (2) cryptic paralogy is an easily identifiable pattern in gene
472 trees. These gene trees typically exhibited polyphyly of individuals representing otherwise
473 monophyletic species and deep divergences between clades, including members of those
474 spuriously polyphyletic species (Fig. S1), a pattern distinct from what is expected via incomplete
475 lineage sorting. This pattern points to a largely artifactual source of cryptic paralogs, as
476 repeated differential loss within species seems an unlikely biological phenomenon. Because our
477 dataset relied heavily on low quality DNA extracted from herbarium tissue, we suspect that
478 differential success in library amplification, missing data, and low coverage in the non-coding
479 regions all contributed to the assembly of erroneously single-copy loci from truly multi-copy
480 genes.

481 As troubling as the prospect of unfiltered paralogs may be for empiricists, cryptic
482 paralogs may be easier to detect than biological pseudo-orthologs (Smith and Hahn 2021b)
483 because they produce a striking, easily identifiable pattern in gene trees (Fig. S1). Our results
484 suggest that recent improvements in HybPiper2, a commonly used pipeline for plant
485 phylogenomics that is actively supported, have taken advantage of this identifiability to improve
486 detection of potential paralogs. The recently updated paralog detection features in HybPiper
487 v.2.0.1 identified that our dataset included a high proportion of stitched contigs in which
488 separate exon contigs were concatenated into a single sequence. This high number of stitched
489 contigs increases the likelihood for chimeric sequences (i.e., contigs stitched together from
490 different gene copies), which may present as cryptic paralogs. Despite the evidence suggesting

that chimeric sequences may be a significant source of gene tree error in our dataset, no genes were flagged by HybPiper v.2.0.1 with a chimera warning. Instead, loci that are consistent with our definition of cryptic paralogs, and that likely represent chimeric sequences, were flagged with paralog by depth warnings. Thus for empiricists analyzing museomic datasets, we suggest examination of the proportion of stitched contigs and that paralog warnings by depth are seriously considered, as the metrics used to identify chimeric sequences may be too strict to effectively identify cryptic paralogs in real world data.

Our *Freziera* target sequence capture dataset is very phylogenetically complex, including significant gene tree discordance, a high proportion of paralogs including many cryptic paralogs, and a confirmed role of introgression in the history of the clade. In the face of the resulting gene tree heterogeneity, the biggest improvements to phylogenetic inference in *Freziera* did not come from filtering gene trees to maximize phylogenetic informativeness; instead, they came from reducing the noise from cryptic paralogs. Our results further suggest that ASTRAL-III is reasonably robust to errant phylogenetic signal from cryptic paralogs: while *bipartition*, *by.eye*, and *hybpiper2.no.warnings* did improve gene tree concordance and support, discordance remained, high gains in support were minimal, and the RF distances between these filtered trees and the consensus topology (i.e., 14, 24, and 22) were similar to the *all.ortholog* phylogeny (Table 2). Despite complexity, relationships among species and subclades of *Freziera* were similar between the unfiltered dataset and the best filtering criteria, with most differences in topology resulting from either a few rogue taxa like *F. magnibracteolata* or from within the Candicans group, which represents a radiation within a radiation.

Curating our dataset via gene tree filtering and the removal of cryptic paralogs resulted in a relatively stable bifurcating species tree that serves as the basis for further modeling of evolutionary processes in *Freziera*. However, even after applying data curation that may inflate support for bifurcating relationships, a strong signal of introgression was identifiable, suggesting that *Freziera*'s phylogenetic history may be characterized by multiple reticulation events. This points to a major challenge in phylogenomics today: our ability and limitations to tease apart specific sources of gene tree discordance in empirical datasets, and to account for these in phylogenetic inference (Morales-Briones et al. 2021; Tricou et al. 2022). It is a significant challenge (perhaps an impossible one) to accurately identify paralogs, pinpoint specific instances of introgression, disentangle incomplete lineage sorting from historical gene flow, and reduce the impact of gene tree estimation error in a single empirical phylogenomic dataset in which all of these sources of discordance are present. To fully investigate evolutionary processes, particularly of recently diverged groups, phylogenetic inference of bifurcating

relationships should be accompanied by the exploration of discordance and sources of reticulate relationships to the extent of our abilities.

CONCLUSION

Freziera has a complex phylogenetic history, characterized by a combination of incomplete lineage sorting, introgression and gene duplication that has resulted in an incredibly high rate of gene tree discordance. Even after filtering cryptic paralogs, which we identify as data artifacts associated with museomic DNA, conflict among gene trees remained high (Table 2; Supplementary Fig. #). This adds to a growing body of literature 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).

Phylogenomic data have the potential to be very complex, both in clades that are well-understood to be recalcitrant like Andean plant radiations (Pease et al. 2016; Vargas et al. 2017), and increasingly in less expected systems like lizards (Barley et al. 2022). To best address this complexity, we advocate for the observation of empirical phylogenomic data, including gene tree alignments and topologies, and that data curation be tailored to unique properties of individual datasets. While automated filtering techniques, assembly parameters, and other automated aspects of phylogenomics are powerful tools to improving phylogenetic inference—as we demonstrated in *Freziera*— they are not a replacement for a deep understanding of a dataset. Data observation allowed us to detect a high proportion of artifactual cryptic paralogs, which we were able to reduce by heeding paralog warnings and filtering gene trees. This data curation was subsequently associated with higher support, lower gene tree conflict, and a more stable species tree— the first for an understudied tropical plant clade that previously lacked any phylogenetic information. While targeted sequence capture of universal loci offers potential, especially for phylogenetic studies relying heavily on DNA from natural history collections, these datasets are not without limitations both related to the nature of the data themselves and to the algorithms we use to process and analyze them. Combining exploration of datasets with deep knowledge of the organismal biology of targeted clades is crucial toward overcoming these limitations and inferring robust phylogenetic hypotheses.

SUPPLEMENTARY MATERIAL

Raw data are available on NCBI's Sequence Read Archive: (*web address added upon acceptance*). Supporting information and processed data available from the Dryad Digital Repository: (*web address added upon acceptance*).

ACKNOWLEDGEMENTS

This research was funded by a Louisiana Board of Regents Research Competitiveness Subprogram grant, the LSU College of Science and Office of Research and Economic Development, and NSF Award DEB-2055525. 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.

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.

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.

Barley A.J., Nieto-Montes de Oca A., Manríquez-Morán N.L., Thomson R.C. 2022. The evolutionary network of whiptail lizards reveals predictable outcomes of hybridization. Science. 377:773–777.

Benjamini Y., Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Series B Stat. Methodol. 57:289–300.

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.

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.

Cai L., Xi Z., Lemmon E.M., Lemmon A.R., Mast A., Buddenhagen C.E., Liu L., Davis C.C. 2021. The Perfect Storm: Gene Tree Estimation Error, Incomplete Lineage Sorting, and Ancient Gene Flow Explain the Most Recalcitrant Ancient Angiosperm Clade, Malpighiales. Syst. Biol. 70:491–507.

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.

Danecek P., Auton A., Abecasis G., Albers C.A., Banks E., DePristo M.A., Handsaker R.E., Lunter G., Marth G.T., Sherry S.T., McVean G., Durbin R., 1000 Genomes Project Analysis Group. 2011. The variant call format and VCFtools. Bioinformatics. 27:2156–2158.

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.

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.

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.

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.

Holder M.T., Anderson J.A., Holloway A.K. 2001. Difficulties in detecting hybridization. Syst. Biol. 50:978–982.

Hughes C.E. 2016. The tropical Andean plant diversity powerhouse. New Phytol. 210:1152–1154.

Hughes C.E., 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.

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.

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.

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.

652 Lagomarsino L.P., Frankel L., Uribe-Convers S., Antonelli A., Muchhala N. 2022. Increased
653 resolution in the face of conflict: phylogenomics of the Neotropical bellflowers
654 (Campanulaceae: Lobelioideae), a rapid plant radiation. Ann. Bot. 129:723–736.

655 Larson D.A., Walker J.F., Vargas O.M., Smith S.A. 2020. A consensus phylogenomic approach
656 highlights paleopolyploid and rapid radiation in the history of Ericales. Am. J. Bot. 107:773–
657 789.

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

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

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

668 Li Z., McKibben M.T.W., Finch G.S., Blischak P.D., Sutherland B.L., Barker M.S. 2021. Patterns
669 and Processes of Diploidization in Land Plants. Annu. Rev. Plant Biol. 72:387–410.

670 MacGuigan D.J., Near T.J. 2019. Phylogenomic Signatures of Ancient Introgression in a Rogue
671 Lineage of Darters (Teleostei: Percidae). Syst. Biol. 68:329–346.

672 Madriñán S., Cortés A.J., Richardson J.E. 2013. Páramo is the world's fastest evolving and
673 coolest biodiversity hotspot. Front. Genet. 4:192.

674 Malinsky M., Matschiner M., Svardal H. 2021. Dsuite - Fast D-statistics and related admixture
675 evidence from VCF files. Mol. Ecol. Resour. 21:584–595.

676 Malinsky M., Svardal H., Tyers A.M., Miska E.A., Genner M.J., Turner G.F., Durbin R. 2018.
677 Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by
678 gene flow. Nat Ecol Evol. 2:1940–1955.

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

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

683 Molloy E.K., Warnow T. 2018. To Include or Not to Include: The Impact of Gene Filtering on
684 Species Tree Estimation Methods. Systematic Biology. 67:285–303.

685 Morales-Briones D.F., Kadereit G., Tefarikis D.T., Moore M.J., Smith S.A., Brockington S.F.,
686 Timoneda A., Yim W.C., Cushman J.C., Yang Y. 2021. Disentangling Sources of Gene
687 Tree Discordance in Phylogenomic Data Sets: Testing Ancient Hybridizations in
688 Amaranthaceae s.l. Syst. Biol. 70:219–235.

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

691 Phytol. 218:1668–1684.

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

694 Nguyen L.-T., Schmidt H.A., von Haeseler A., Minh B.Q. 2015. IQ-TREE: a fast and effective
695 stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol.
696 32:268–274.

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

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

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

703 Pease J.B., Haak D.C., Hahn M.W., Moyle L.C. 2016. Phylogenomics Reveals Three Sources
704 of Adaptive Variation during a Rapid Radiation. PLoS Biol. 14:e1002379.

705 Puritz J.B., Hollenbeck C.M., Gold J.R. 2014. dDocent: a RADseq, variant-calling pipeline
706 designed for population genomics of non-model organisms. PeerJ. 2:e431.

707 Rambaut. 2014. FigTree v1. 4.2, a graphical viewer of phylogenetic trees. //tree. bio. ed. ac.
708 uk/software/figtree/right angle bracket.

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

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

714 Santamaría-Aguilar D., Monro A.K. 2019. Compendium of *Freziera* (Pentaphylacaceae) of
715 South America including eleven new species and the typification of 22 names. Kew Bull.
716 74:14.

717 Särkinen T., Staats M., Richardson J.E., Cowan R.S., Bakker F.T. 2012. How to open the
718 treasure chest? Optimising DNA extraction from herbarium specimens. PLoS One.
719 7:e43808.

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

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

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

727 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.

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

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.

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.

Tricou T., Tannier E., de Vienne D.M. 2022. Ghost Lineages Highly Influence the Interpretation of Introgression Tests. Syst. Biol. 71:1147–1158.

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 Ulloa C., Acevedo-Rodríguez P., Beck S., Belgrano M.J., Bernal R., Berry P.E., Brako L., Celis M., Davidse G., Forzza R.C., Gradstein S.R., Hokche O., León B., León-Yánez S., Magill R.E., Neill D.A., Nee M., Raven P.H., Stimmel H., Strong M.T., Villaseñor J.L., Zarucchi J.L., Zuloaga F.O., Jørgensen P.M. 2017. An integrated assessment of the vascular plant species of the Americas. Science. 358:1614–1617.

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.

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.

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.

Figure and Table Captions

Figure 1. Morphological diversity and geographic distribution of Freziera. Photos on the left illustrate their habit as shrubs (a), leaf variation (b-f), typical flower (b, c) and fruit (d, e) morphology, and the characteristic apical bud (f). Map on the right illustrates distribution, with highest density in montane regions of the Neotropics. Photos by L. Lagomarsino.

Figure 2. Impact of gene filtering of criteria on gene tree discordance. (a) Venn diagram showing overlap in genes selected by the three top-performing filtering metrics: bipartition support, by eye, and loci without any paralog warnings from HybPiper v.2.0.1. The total number of genes included in each dataset is given in parentheses below the respective dataset name. (b) Phyparts measurements of discordance between gene trees and major nodes of the species tree for (clockwise from the top left) bipartition, by.eye, hybpiper2.no.warnings, and all.orthologs datasets. Clade position and color correspond to the consensus topology in Fig. 3.

Figure 3. Species tree topology of *Freziera*, including (a) branch lengths in coalescent units for the *bipartition* ASTRAL-III tree, (b) the consensus topology for all 14 datasets with bootstrap support for each node and named clades labeled at tips, and (c) cartoon trees of named clades depicting the consensus topology and the most frequent deviations from the consensus (see Table 2 for a comparison of the species tree for each dataset to the consensus tree).

Table 1. Names applied to each of the 14 datasets comprised of HybPiper v.1.3.1 gene trees, descriptions of the criteria used and thresholds set for gene or gene tree filtering, and the number of loci selected by each metric..

Table 2. Summary of results for the 16 ASTRAL analyses of datasets following different filtering criteria. Dataset names follow those defined in Table 1, except for two datasets that followed the same filtering criteria but were comprised of HybPiper v.2.0.1 gene trees (h2gt; indented). 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 each species tree relative to the consensus topology. Values for the four best-performing datasets are bolded for each metric. The four columns on the right summarize major topological conflicts (illustrated in Fig. 3c) 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 mark disagreement from the alternative placement of only one species.

Appendix 1. Supplementary Materials and Methods

Figure S1. Gene trees and alignments demonstrating *by.eye* determination for genes. Genes were thought be single-copy if gene trees exhibited a relatively shallow backbone and some clade structure in common with the species tree (a, b); typically, the placement and monophyly of the Humiriifolia group (pink) was a good indicator (note the non-monophyly in c and d). Additionally, alignments for these genes exhibited only several variable sites in exons and

characters at variable sites were, for the most part, consistent within species (e, f). Genes were thought to be “multi-copy”—which includes the presence of multiple sequences due to allelic variation as well as gene/genome duplication—if relatively long internal branches were observed along with the non-monophyly of species or clades (c, d). In instances were a clade comprising individuals from species distributed across the species tree with a long internal branch nested within *Freziera* (c), it was suspected that the observed topology was the result of allelic variation: sequences that were mostly similar to others in the alignment but that clustered together due to a few key differences. These differences were typically observable in alignments, and, notably, sequences differed within species (e.g., *F. callophyla*, *F. candicans*, and *F. chrysophylla* in g). Finally, genes were suspected to by cryptic paralogs if the gene tree was composed of two or more deeply-divergent clades in which aspects of the species tree was recapitulated in those clades (e.g., members of the Humiriifolia clade being non-monophyletic but forming sister groups to the remaining members of their respective clades, as in d). Alignments exhibited sequence “motifs” where there were a relatively high number of variable sites in exons, again, samples within species belonged to different motifs (e.g., *F. alata*, *F. incana*, *F. lanata*, *F. magnibracteolata*, and *F. yanachagensis* in h). While patterns displayed here for multi-copy alignments (g, h) are largely similar, a key difference is that the supposed allelic variation in genes similar to g was often localized to a single exon. The motifs of supposed cryptic paralogs (h) were present in multiple exons and often the entire alignment. Additionally, groupings of samples belonging to motif 1 and motif 2 of one exon were not the same and groupings of samples belonging to motifs observed in the next exon of the same gene. This suggested that the variation observed may be the result of chimeric sequences produced by stitching contigs of an undetected multi-copy gene (i.e., a cryptic paralog).

Figure S2. ASTRAL-III species trees for the 16 filtered datasets: (a) *clocklike.bipartition*, (b) *tree.length*, (c) *PI.per.branch*, (d) *high.percent.internal*, (e) *proportion.PI*, (f) *average.BS*, (g) *low.percent.internal*, (h) *bipartition*, (i) *variable.sites*, (j) *by.eye*, (k) *hybpiper2.no.warnings*, (l) *h2.no.warnings.h2gt*, (m) *hybpiper2.orthologs*, (n) *h2.orthologs.h2gt*, (o) *1000bp*, and (p) *all.orthologs*.

Figure S3. Phyparts summary of conflict and concordance among gene trees and the species tree for the (a) *all.orthologs*, (b) *bipartition.support*, (c) *by.eye*, and (d) *hybpiper2.no.warnings* datasets. 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 dominant alternative for that clade (green), all other alternatives (red), and the proportion that have less than 50% bootstrap support (grey).

Figure S4. Results of D and f4 statistics calculated with Dsuite on the “bipartition support”, “by eye”, and “HybPiper2 no warnings" datasets. The columns indicate 1) values of Patterson’s D-statistic for evaluated trios, 2) highly significant p-values associated with the D-statistics after Benjamini-Hochberg correction, and 3) f4 ratio statistics. Introgression is detected across all datasets even after assessing family-wise error rate.

Table S1. 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/), collection date, provenance, and sample ID used in phylogenetic analyses--and summary statistics generated with HybPiper. Detailed

849 descriptions of these columns are available at
850 (https://github.com/mossmatters/HybPiper/wiki).

851 **Table S2.** Per locus statistics for alignments including outgroup sequences (columns A-
852 V) and excluding outgroup sequences (columns W-AN). Root-to-tip variation, tree
853 length, bipartition support, average bootstrap values, and percent of internal branch
854 lengths in the total tree length (columns B-F) were calculated from gene trees.

855 **Table S3.** Summary of the proportion of stitched contigs per locus and paralog warning
856 by depth based on the updated statistics from HybPiper v.2.0.1. By eye determination
857 (see Fig. S1 for details), long paralog warnings from HybPiper v.2.0.1, and paralog
858 warnings from HybPiper v.1.3.1 are also indicated for loci.

859 **Table S4.** Gene stats and gene lengths for threshold tests using *F. candicans*.