Running title: Moore-Pollard et al. – Compositae-ParaLoss-1272 probe set

Compositae-ParaLoss-1272: Complementary sunflower specific probe-set reduces paralogs in phylogenomic analyses of complex systems

Erika R. Moore-Pollard[1], Daniel S. Jones[2], Jennifer R. Mandel[1]

1 Department of Biological Sciences, University of Memphis, 3700 Walker Ave, Memphis, Tennessee 38152, U.S.A.

2 Department of Biological Sciences, Auburn University, 101 Rouse Life Sciences, Auburn, Alabama 36849, U.S.A.

Manuscript received ___; Revision accepted ___

Number of words: 5,341 words

Emails of all authors:

Erika R. Moore-Pollard*, moore.erika.r@gmail.com

Jennifer R. Mandel*, jmandel@memphis.edu

Daniel S. Jones, danielsjones@auburn.edu

* = *Addresses for correspondence*

ABSTRACT

Premise: The sunflower family-specific probe set, Compositae-1061, enables family-wide phylogenomic studies and investigations at lower-taxonomic levels, but may lack resolution at genus to species levels, especially in groups complicated by polyploidy and hybridization.

Methods: We developed a Hyb-Seq probe set, Compositae-ParaLoss-1272, which targets orthologous loci in Asteraceae. We tested its efficiency across the family by simulating target-enrichment sequencing in silico. Additionally, we tested its effectiveness at lower taxonomic levels in the historically complex genus *Packera*. We performed Hyb-Seq with Compositae-ParaLoss-1272 for 19 *Packera* taxa which were previously studied using Compositae-1061. Sequences from both probe sets, plus a combination of both, were used to generate phylogenies, compare topologies, and assess node support.

Results: We report that Compositae-ParaLoss-1272 captured loci across all tested Asteraceae members, had less gene tree discordance, and retained longer loci than Compositae-1061. Most notably, Compositae-ParaLoss-1272 recovered substantially less paralogous sequences than Compositae-1061, with only ~5% of the recovered loci reporting as paralogous, compared to ~59% with Compositae-1061.

Discussion: Given the complexity of plant evolutionary histories, assigning orthology for phylogenomic analyses will continue to be challenging. However, we anticipate Compositae-ParaLoss-1272 will provide improved resolution and utility for studies of complex groups and lower-taxonomic levels in the sunflower family.

Keywords: Asteraceae; double-capture; Hyb-Seq; MarkerMiner; *Packera*; polyploidy; Target-Enrichment

INTRODUCTION

The sunflower family, also known as the daisy family, Asteraceae, or Compositae, is one of the largest flowering plant families making up roughly 10% of all angiosperms. This large and diverse group has presented many challenges for resolving evolutionary relationships and studying diversifications through time and space. Recent phylogenetic work in the family has employed various methods to reconstruct family-level phylogenies to better understand the evolutionary history and relationships of Asteraceae. For example, Huang et al. (2016) used transcriptome data, Zhang et al. (2021) used a combination of transcriptome and whole-genome sequence data, while Mandel et al. (2019) used Target-Enrichment sequencing with a custom probe set designed to enrich for conserved gene sequences in Asteraceae (Mandel et al., 2014, 2017). This probe set has become popular among researchers studying members of Asteraceae and has enabled investigations at lower taxonomic levels, especially understudied groups (e.g., Lichter-Marck et al., 2020; Thapa et al., 2020; de Lima Ferreira et al., 2022; Siniscalchi et al., 2019, 2023).

Targeted sequence probe sets have grown in popularity over the last 10 years with sets designed to target loci across large plant groups: bryophytes (i.e., mosses; Liu et al., 2019), pteridophytes (i.e., ferns, Wolf et al., 2018), and angiosperms (i.e., Johnson et al., 2019), as well as for specific plant families (i.e., Asteraceae, Mandel et al., 2014, 2017; Fabaceae, Chapman, 2015; Ochnaceae, Shah et al., 2021; Orchidaceae, Eserman et al., 2021). Typically, low-coverage genome-skim and/or transcriptome data have been used to design probe sets (Straub et al., 2012; Weitemier et al., 2014; Folk et al., 2015; Fonseca and Lohmann, 2020); however, genome-skimming is generally not as effective for designing a probe set for nuclear genes, as low-coverage genome skim data typically enriches for organellar genomes and other high-copy

genomic sequences in plants (Stull et al., 2013). These genomic regions are often highly conserved and repetitive and are thus less useful for resolving relationships in some groups. Using transcriptome data offers the potential to sequence and select from thousands of loci, enabling the survey of genomic regions with different rates of molecular evolution.

Several tools have recently become available to design targeted sequence probe sets using transcriptome data more easily, such as OrthoFinder (Emms and Kelly, 2019) and MarkerMiner (Chamala et al., 2015). OrthoFinder is a pipeline that identifies orthogroups and/or orthologs in transcriptomes based on sequence similarities across many species (Emms and Kelly, 2015). In return, the output returns a list of exons usable for probe design. One disadvantage to OrthoFinder, and ultimately the transcriptome-only approach, is that without knowledge of intron-exon topology, probes could overlap boundaries and thus would not be effective at sequence capture (McKain et al., 2018). Alternatively, identification of intron-exon boundaries is straightforward in the MarkerMiner tool, which aligns transcriptome data to reference angiosperm genome sequences and returns intron-masked multiple sequence alignments (Chamala et al., 2015; McKain et al., 2018). The general workflow for MarkerMiner compares user-provided transcriptome sequences against reference genomes with known single-copy orthologous genes (e.g., *Arabidopsis thaliana* (L.) Heynh.), drastically reducing the number of paralogous sequences, or 'paralogs', retained for each gene. Probe sets designed using this approach have yielded greater phylogenetic resolution in some groups at the family level (e.g., Cactaceae; Acha and Majure, 2022) and genus/species level (e.g., *Euphorbia* L.; Villaverde et al., 2018; *Zanthoxylum* L., Reichelt et al., 2021). Retaining only single-copy orthologs as a result of MarkerMiner can greatly improve species tree inference as paralogs complicate phylogeny building by causing gene tree heterogeneity. If not accounted for properly, this

heterogeneity can lead to misleading phylogeny construction and an incorrect interpretation of species relationships (Smith and Hahn, 2021).

In this study, we used 48 transcriptomes to generate a new probe set for sequencing orthologous sequences in Asteraceae utilizing MarkerMiner. Our sampling included 45 Asteraceae taxa and three outgroups from across the order Asterales: Calyceraceae, Campanulaceae, and Goodeniaceae. Though Compositae-1061 has been shown to be efficient at higher- and some lower-taxonomic levels within the family, it generally lacks resolution at the genus to species level. Therefore, we designed this probe set with the aim to provide higher resolution at lower-taxonomic levels and help tackle challenges associated with paralogy, especially among complex groups. To do this, we tested the compatibility and efficiency of this new probe set across the entire family by simulating target-enrichment sequencing in silico in six Compositae members spanning across the family. We then used members of the genus *Packera* Á. Löve & D. Löve as a model system to directly test the efficacy of the probe set by sequencing 16 *Packera* and three outgroup taxa using this newly designed probe set, named Compositae-ParaLoss-1272, and the Compositae-1061 probe set. Additionally, we combined the Compositae-1061 and Compositae-ParaLoss-1272 sequence data to represent an in silico double-capture method. We then generated phylogenetic trees, compared their topologies, and assessed node support to determine whether Compositae-ParaLoss-1272 provided greater resolution at the genus/species level compared to Compositae-1061.

METHODS

Probe Development

To identify single-copy nuclear loci and select regions for target enrichment probe design, transcriptome data from 48 taxa spanning Asterales were compiled from the 1KP

initiative (One Thousand Plant Transcriptomes Initiative, 2019), Sunflower Genome database (https://sunflowergenome.org/), or generated de novo (Appendix S1; see Supporting Information with this article). Four specimens were collected from the Memphis Botanic Garden live collection, of which we did not make an herbarium voucher. All 48 samples were used as input for MarkerMiner v. 1.0 (Chamala et al., 2015) using default settings with both *Arabidopsis thaliana* and *Vitis vinifera* L. as reference genomes. MarkerMiner is an open access, bioinformatic workflow that compares user-provided transcriptomes against reference angiosperm genomes with known single-copy orthologous genes that can be used to design primers or probes for targeted sequencing. Orthologous genes are classified as single copy in the reference genomes if they are present across 17 genomes that were previously annotated as part of a systematic survey on duplication resistant genes (De Smet et al., 2013). We aimed for this new probe set to have no gene overlap with Compositae-1061 (Mandel et al., 2014, 2017) and Angiosperm-353 (Johnson et al., 2019). Therefore, if a gene present in our new probe set was in either Compositae-1061 or Angiosperm-353, we removed it from our targeted gene list, e.g., if AT3G47610 was included in the Angiosperm-353 gene list and ours, we removed this gene from our list and did not design probes for it.

Exons with lengths ranging from 120 - 1,000bp and a minimum variability of two single nucleotide polymorphisms (SNPs) were selected using a custom python script (https://github.com/ClaudiaPaetzold/MarkerMinerFilter). The resulting 3,853 exonic regions, spanning 1,925 genes around 1,112 - 85,780bp long (Appendix S2), were further processed by MyBaits at Arbor Biosciences (Ann Arbor, Michigan, USA) to produce a set of 120-mer tiled baits that overlap every 60 bases and share an 80% identity when possible, similar to methods used to develop the MyBaits Compositae-1061 kit (Mandel et al., 2014), hereafter referred to as

Comp-1061. Additional filtering steps were implemented as follows: 1) sequence clusters containing five or more taxa not targeting lineage specific genes or clusters were retained, 2) clusters containing only the reference sequence data were removed, 3) probes with at least three sequences that covered the alignment were retained, and 4) probes with high similarities (80% or 90%) representing only one or two species were collapsed. Finally, two additional loci were added to the probe design: the MADS-box transcription factor *LEAFY* (*LFY*, Weigel et al. 1992) and the transmembrane pseudokinase *CORYNE* (*CRN*, Müller et al., 2008), two conserved single-copy genes that regulate flower development and meristem size, respectively, in Angiosperms. Gene sequences for *LFY* were identified using the tblastx plugin in Geneious Prime v. 2023.0.4 (https://www.geneious.com) with custom *Bidens ferulifolia* (Jacq.) Sweet (cv. Compact Yellow) leaf transcriptome and *Lactuca sativa* L. genome assembly (v.8) blast databases respectively. The *CRN* gene sequence (AT5G13290) came directly from *Arabidopsis thaliana* using The Arabidopsis Information Resource (TAIR, https://www.arabidopsis.org/).

The resulting MyBaits target enrichment kit contains 60,158 120bp-long, in-solution, biotinylated baits based on target sequence information. The final bait panel, Compositae-ParaLoss-1272, consisted of 13,117 probes and 1,272 loci after filtering (Table 1).

These methods are compared to Comp-1061, which was developed via BLAST searches of expressed-sequence tag (EST) data from three species within the sunflower family (*Helianthus annuus* L. [sunflower], *Lactuca sativa* [lettuce], and *Carthamus tinctorius* L. [safflower]) to a set of previously identified *Arabidopsis thaliana* single-copy genes. This resulted in 1,061 genes, for which 9,678 biotinylated baits were designed (Mandel et al., 2014, 2017). Refer to Table 1 for a comparison between Compositae-ParaLoss-1272 and Comp-1061.

Simulating capture sequencing across Compositae

We simulated a target-enrichment sequencing run in silico on six published genomes spanning Asteraceae (Figure 1) using Compositae-ParaLoss-1272, hereafter referred to as Comp-ParaLoss-1272, and Comp-1061 in the software CapSim (Cao et al., 2018) to investigate the efficiency of this new probe set for recovering loci across the sunflower family. CapSim is a tool that simulates a sequence run in silico with given a genome sequence and probe set as input. The simulated data can be used for evaluating the performance of the analysis pipeline, as well as the efficiency of the probe design.

Prior to running CapSim, an index file was generated, and probes were aligned to the six genomes using Bowtie2 v. 2.3.5.1 (Langmead and Salzberg, 2012; Langmead et al., 2019). After the alignment, the sam files were sorted and indexed into bam files using samtools v. 1.9 (Danecek et al., 2021). The resulting bam files were then used as input in CapSim using the *jsa.sim.capsim* command with the following settings: median fragment size at shearing (--fmedian) set to 250, miseq simulated (--miseq), illumina read length (--illen) set to 150, and the number of fragments (--num) set to 50,000,000. The resulting fastq files were used as input in the HybPiper v. 2.0.1 (Johnson et al., 2016) pipeline to map simulated sequences against the probe set. Summary and paralog statistics were recovered using the 'stats' and 'paralog_retriever' options in HybPiper.

Specimen collection

An Illumina sequence run was performed using the new probe set on a selection of 19 total taxa, 16 *Packera* and three outgroup taxa, that were previously sequenced with the Comp-1061 probe set (Moore-Pollard and Mandel, 2023a). *Packera* taxa were selected to be representative across the entire *Packera* phylogenetic tree from Moore-Pollard and Mandel (2023a). One outgroup taxon, *Packera loratifolia* (Greenm.) W.A.Weber & Á.Löve, was

included in this analysis as an outgroup instead of an ingroup since previous studies have shown it is likely misclassified in *Packera* and instead should be in *Senecio* (Barkley, 1985; Bain and Jansen, 1995; Bain and Golden, 2000; Pelser et al., 2007; Moore-Pollard and Mandel, 2023a). A complete list of sampled species, herbarium vouchers, and NCBI accession numbers can be found in Table 2.

DNA extraction and sequencing

DNA extraction and sequencing methods for the 19 taxa utilizing the Comp-ParaLoss-1272 probe set followed steps outlined by Moore-Pollard and Mandel (2023a). Briefly, dried leaf tissue collected from herbarium specimens was used to extract DNA. DNA length was assessed by running a 1% agarose gel in 1X TBE and GelRed 3x (Biotium), with a target DNA length of 400-500 base pairs (bp). If DNA fragments appeared larger than 500bp, up to 1µg DNA was sheared via sonication with a QSonica machine (amp: 20%; pulse: 10 seconds on, 10 seconds off) (ThermoCube, New York, USA). Sheared DNA was then used to generate barcoded libraries utilizing NEBNext Ultra II DNA Library Prep Kit (New England Biolabs, Ipswich, Massachusetts, USA). Libraries produced followed the NEBNext Ultra II Version 5 protocol with size selection on DNA fragments at 300-400bp range but were adjusted by halving the amount of reagents and DNA. Targeted sequence capture was performed on the libraries using the newly designed probe set, Comp-ParaLoss-1272, from Arbor Biosciences (Ann Arbor, Michigan, USA) described above, following manufacturer's protocols (version 4.01). Captured targets were amplified and quantified using KAPA library quantification kits (Kapa Biosystems, Wilmington, Massachusetts, USA). Quality and quantity checks were performed throughout using a Nanodrop 2000 (Thermo Fisher Scientific, Carlsbad, California, USA) and Qubit High Sensitivity assay (ThermoFisher Scientific, Oregon, USA), respectively. The pooled libraries

208 were sequenced on an Illumina NovaSeq6000 at HudsonAlpha Institute of Technology
209 (Huntsville, Alabama, USA). Data for the Comp-1061 taxa were obtained from Moore-Pollard
210 and Mandel (2023a) and available at NCBI (Bioproject: PRJNA907383).

211 **Phylogenetic analyses**

212 Raw sequence reads from Comp-1061 and Comp-ParaLoss-1272 were cleaned and
213 trimmed of adapters using Trimmomatic v. 0.36 (Bolger et al., 2014), implementing the Sliding
214 Window quality filter (illuminaclip 2:30:10, leading 20, trailing 20, sliding window 5:20).
215 Cleaned reads were retained if they had a minimum length of 36 bp. Cleaned reads were then
216 mapped against the corresponding loci targeted in the Comp-1061 (Mandel et al., 2014) or
217 Comp-ParaLoss-1272 probe sets using the HybPiper pipeline. A combined reference/de novo
218 assembly was performed using BWA v. 0.7.17 (Li and Durbin, 2009) and SPAdes v. 3.5
219 (Bankevich et al., 2012), respectively, with specified kmer lengths: 21, 33, 55, 77, and 99.
220 Resulting sequences were then aligned using MAFFT v. 7.407 (Katoh and Standley, 2013).
221 Maximum likelihood trees were built in RAxML v. 8.1.3 (Stamatakis, 2014) with 1,000
222 bootstrap replicates under the GTR+I+Γ model. Species trees were generated from each
223 resulting RAxML gene matrix using ASTRAL-III v. 5.7.3 (Zhang et al., 2018), a pseudo-
224 coalescent tree building method. Local posterior probability (LPP) values were generated at each
225 node to indicate the probability that the resulting branch is the true branch given the set of input
226 gene trees. LPP is considered a more reliable clade support measure than bootstrapping since it is
227 computed based on a quartet score (Sayyari and Mirarab, 2016) and assumes incomplete lineage
228 sorting (Zhang et al., 2018).

229 The sequence data from Comp-1061 and Comp-ParaLoss-1272 were also combined,
230 hereafter referred to as Comp-1061 + Comp-ParaLoss-1272, and a phylogenetic tree was built

following the methods above. The resulting species trees, Comp-1061, Comp-ParaLoss-1272, and Comp-1061 + Comp-ParaLoss-1272 were then visualized using the package *phytools* (Revell, 2012) in R v. 4.0.5 (R Core Team, 2016; RStudio, 2020).

Measuring phylogenomic discordance

To determine if Comp-ParaLoss-1272 increased node resolution across *Packera*, Quartet Sampling (Pease et al., 2018) was used to assess the confidence, consistency, and informativeness of internal tree relationships. Quartet Sampling provides a more comprehensive support value estimate than LPP by calculating four scores, three at each node (quartet concordance [QC], quartet differential [QD], and quartet informativeness [QI]) and one at the tip (quartet fidelity [QF]), to determine if the internal relationships are caused by a lack of data, underlying biological processes, or rogue taxa. QC specifies how often a concordant quartet is inferred over other discordant quartets as a range from -1 to 1: -1 indicates that the quartets are more often discordant than concordant and 1 indicates that all quartets are concordant. QD reveals how skewed the discordant quartets are as a range from 0 (high skew) to 1 (low skew). QI suggests how informative the quartets are as a range from 0 (none are informative) to 1 (all are informative). Each terminal branch is then given a QF score which reports how often a taxon is included in the concordant topology given a range of 0 (taxon is present in none) to 1 (taxon is present in all). Quartet Sampling requires a concatenated nucleotide matrix and a rooted species tree. The concatenated matrices were generated using FASconCAT-G v. 1.02 (Kück and Longo, 2014) into a phylip format. The input phylogeny was then rooted using the *pxrr* command in Phyx (Brown et al., 2017).

PhyParts v. 0.0.1 (Smith et al., 2015) was then used to quantify and visualize discordance in the final phylogenies. PhyParts summarizes and visualizes conflict among gene trees given the

resulting species tree topology by performing a bipartition analysis, which helps determine if the node support values are misleading because of underlying discordance. This tool requires a rooted final species tree and rooted gene trees as input. Thus, these trees were rooted to the three outgroup taxa, *Roldana gilgii* (Greenm.) H.Rob. & Brettell, *Emilia fosbergii* Nicolson, and *Packera loratifolia*. The script "phypartspiecharts.py" (available at https://github.com/mossmatters/MJPythonNotebooks) was then used to map pie charts onto the nodes in the final species tree, detailing whether there is one dominant topology in the gene trees with not much conflict, if there is one frequent alternative topology, or many low-frequency topologies.

To estimate similarity scores between the Comp-1061 and Comp-ParaLoss-1272 tree topologies, we calculated the adjusted Robinson-Foulds (RF_{adj}) distance as outlined by Moore-Pollard and Mandel (2023a) between the two trees using the RF.dist function in package *phangorn* (Schliep, 2011) in R. Unrooted ASTRAL-III trees were used as input with the "normalize" argument set to TRUE. RF_{adj} calculates the distance between two unrooted trees, with resulting RF_{adj} values closer to zero indicating that the tree topologies are similar, and values closer to one show complete dissimilarity. Parsimony informativeness was calculated between matrices of Comp-1061 and Comp-ParaLoss-1272 using MEGA-X: Molecular Evolutionary Genetics Analysis across computing platforms v. 10.2.5 (Kumar et al., 2018). Heatmaps to compare sequence lengths of retained loci between probe sets were generated in R using the package *ggplot2* (Wickham, 2016). Additionally, the average and standard deviation of locus lengths were calculated using the mean and sd functions in base R.

RESULTS

CapSim

CapSim results showed that both the Comp-1061 and Comp-ParaLoss-1272 probe sets were successful across a broad range of Asteraceae members since both probe sets retained a moderate number of loci. The Comp-1061 probe set generally retained more loci than Comp-ParaLoss-1272 with an average of about 551 loci retained using the Comp-1061 probe set, and an average of 453 loci with the Comp-ParaLoss-1272 probe set (Table 3). Even so, the average length of the loci was much longer in the Comp-ParaLoss-1272 probe set with genes averaging 1,922bp long, and the Comp-1061 probe set produced genes averaging 403bp long (Appendix S3). Additionally, Comp-ParaLoss-1272 produced fewer paralog warnings than Comp-1061 with a range of 0-2 paralogs retained per sample with the Comp-ParaLoss-1272 probe set, and a range of 96-250 paralogs per sample with Comp-1061 (Table 3). A full list of statistics can be found in Appendix S3.

Packera sequence stats

Illumina sequencing utilizing the Comp-ParaLoss-1272 probe set resulted in a total of 501 million reads and 76 billion sequences across the 19 newly sequenced taxa. Additionally, the minimum and maximum number of reads ranged from 10.4 million in *Emilia fosbergii* to 90.1 million in *Packera streptanthifolia* (Greene) W.A.Weber & Á.Löve. (Table 2). The Comp-1061 sequence data from Moore-Pollard & Mandel (2023) totaled 142 million reads and 21 billion sequences, with the minimum and maximum number of reads ranging from 1.2 million in *Packera musiniensis* (S.L.Welsh) Trock to 15 million in *Packera dubia* (Spreng.) Trock & Mabb., respectively.

The HybPiper pipeline retained 1,049 genes (out of 1,061) when using the Comp-1061 probe set, and 1,213 genes (out of 1,272) with the Comp-ParaLoss-1272 probe set. The number of loci recovered for each taxon ranged from 923 in *Packera musiniensis* to 1,051 in *Roldana*

gilgii using the Comp-1061 probe set, and 1,258 in *Packera musiniensis* to 1,271 in *Packera streptanthifolia* using the Comp-ParaLoss-1272 probe set. The number of loci retained was proportionally higher in Comp-ParaLoss-1272 compared to Comp-1061 (Figure 2B), though the Comp-1061 alignment contained fewer missing data (Comp-1061: 34.89%; Comp-ParaLoss-1272: 35.05%) and was more parsimony informative (Comp-1061: 11.7%; Comp-ParaLoss-1272: 8.3%) than Comp-ParaLoss-1272 (Appendix S6). Alternatively, the Comp-ParaLoss-1272 probe set recovered drastically fewer paralogous sequences ('paralogs') than the Comp-1061 probe set, with only about 5% of the recovered loci reporting as paralogous, compared to 59% with the Comp-1061 probe set (Figure 2A). The number of paralog warnings ranged from 35-407 genes per sample with the Comp-1061 probe set, compared to 0-14 in the Comp-ParaLoss-1272 probe set (Table 4). Additionally, Comp-ParaLoss-1272 recovered much longer loci compared to Comp-1061 ($Mean_{Comp\text{-}1061} = 292.13$, $SD_{Comp\text{-}1061} = 146.18$; $Mean_{Comp\text{-}ParaLoss\text{-}1272} = 1{,}192.02$, $SD_{Comp\text{-}ParaLoss\text{-}1272} = 809.5$; Figure 3). Combining the probe sets, Comp-1061 + Comp-ParaLoss-1272, resulted in a species tree made from 2,182 loci (out of 2,333). Refer to Appendix S6 for a full compilation of statistics.

Discordance of *Packera* taxa

A higher number of gene trees were represented in the final Comp-ParaLoss-1272 species tree compared to the Comp-1061 tree (Normalized quartet score = 0.461 and 0.424, respectively), with the Comp-1061 + Comp-ParaLoss-1272 species tree having an intermediate value (Normalized quartet score = 0.436). Additionally, the Comp-ParaLoss-1272 probe set provided higher resolution at internal nodes compared to the previous probe set, with 13 of the 17 internal nodes having local posterior probability (LPP) values greater than or equal to 0.97LPP, eight of those being fully supported (1.0LPP). This is compared to the Comp-1061

probe set which only had eight nodes greater than or equal to 0.97LPP, seven of those with 1.0LPP (Figure 4), while Comp-1061 + Comp-ParaLoss-1272 had 12 nodes greater than or equal to 0.97 LPP, nine of which were 1.0LPP (Appendix S4). Additionally, the level of discordance of internal *Packera* relationships varied between both trees. Quartets are more often discordant than concordant in the Comp-1061 tree, with four internal nodes having negative Quartet Concordance (QC) values, compared to only one node (between *Packera pseudaurea* (Rydb.) W.A.Weber & Á.Löve and *P. aurea* (L.) Á.Löve & D.Löve, QC = -0.3) in the Comp-ParaLoss-1272 tree (Figure 5).

The resulting Comp-1061 and Comp-ParaLoss-1272 species tree topologies were moderately incongruent with each other (RF_{adj} = 0.625). Of the taxon relationships that remained the same in both trees, Comp-ParaLoss-1272 showed more concordant and strongly supported relationships compared to Comp-1061 (Figures 5 and 6). For example, both tree topologies have *P. cynthioides* (Greene) W.A.Weber & Á.Löve and *P. candidissima* (Greene) W.A.Weber & Á.Löve as sister, and *P. franciscana* (Greene) W.A.Weber & Á.Löve and *P. texensis* O'Kennon & Trock as sister; all four within the same smaller clade (Figure 5). However, the node between *P. franciscana* and *P. texensis* and the node joining the two sister groups were majorly discordant in the Comp-1061 tree (QC = -0.0032, -0.32; respectively), while the same relationships in the Comp-ParaLoss-1272 tree were less discordant (QC = 0.16, 0.078; respectively). Even so, the internal relationships were still not strongly supported.

The outgroup relationships and monophyly of *Packera* were fully supported in the Comp-ParaLoss-1272 tree (Figure 5). Alternatively, the Comp-1061 tree showed the monophyly of *Packera* with full support; however, the relationship between the outgroup taxa, *Emilia fosbergii* and *Roldana gilgii,* showed weak support with a discordant skew (QS score at node: 0.3/0/1;

Figure 5). Quartet fidelity (QF) scores were generally higher in the Comp-ParaLoss-1272 tree than the Comp-1061 tree, which ranged from 0.57-0.79 and 0.42-0.64, respectively (Figure 5), indicating a higher percentage of quartet topologies involving the tested taxa were concordant with the focal tree branch in the Comp-ParaLoss-1272 tree.

DISCUSSION

In this study, we designed and tested a complementary Compositae-specific probe set, Compositae-ParaLoss-1272, that provided higher resolution at the lower-taxonomic levels of species in our *Packera* test case. The new probe set dramatically reduced the number of paralogs recovered, retained longer gene sequences, and was likely important for improving the resolution in our *Packera* comparison. Also, this new probe set successfully retained genes across all tested members of Asteraceae and recovered more and longer orthologous genes than Comp-1061 (Appendix S3), as well as retained a substantially lower number of paralogs than Comp-1061 (Table 3) when tested in silico. Finally, there is the ability to do a double sequence capture since the genes associated with Comp-1061 and Angiosperm-353 are not included in the Comp-ParaLoss-1272 probe design (Table 1).

While our results showed that Comp-1061 retained a higher number of genes in silico (Table 3), the Illumina sequencing run of the Comp-ParaLoss-1272 probe set shows much higher locus retention and greater resolution than the Comp-1061 probe set (Table 3). We hypothesize that the low loci retention in silico is a relic of read simulators not always capturing the variances of Illumina sequenced data since they cannot model noise or sequencing technology biases perfectly (May et al., 2022; Duncavage et al., 2023). Additionally, we suspect that having longer gene sequences in the probe set influences read simulator results, though we cannot confirm the validity of these suspicions.

Comp-ParaLoss-1272 contained more missing data and was considered slightly less parsimony informative (PI) than Comp-1061 (Appendix S6); however, the differences were minimal (PI_{1272} = 23.4%, PI_{1061} = 24.1%). Interestingly, similar results were found in a previous study that generated a Fabaceae specific probe set using MarkerMiner and compared the results to other probe design methods (Vatanparast et al., 2018). This study found that MarkerMiner produced fewer paralogous loci than other design methods, but also was not as parsimony informative as other methods, following our results.

When comparing the Comp-1061 and Comp-ParaLoss-1272 tree topologies to the larger *Packera* phylogeny (Moore-Pollard and Mandel, 2023a), the Comp-ParaLoss-1272 tree's evolutionary relationships was in slightly higher agreement with the whole-genus phylogeny (RF_{adj} = 0.6) as compared to Comp-1061 (RF_{adj} = 0.667) (Appendix S5), potentially indicating this new probe set is more robust to species sampling compared to Comp-1061. For example, our Comp-1061 tree places *P. layneae* (Greene) W.A.Weber & Á.Löve as sister to the remaining core *Packera* species. This relationship differs from both the Comp-ParaLoss-1272 and Moore-Pollard and Mandel (2023a) trees, which have *P. layneae* placed more deeply nested and with other Californian endemic species (Figure 4; Moore-Pollard and Mandel, 2023a). Additionally, the placement of *P. glabella* (Poir.) C.Jeffrey in the Comp-1061 tree differs from past phylogenomic studies, including the Comp-ParaLoss-1272 tree in this study, which place it as sister to all remaining *Packera* taxa (Freeman, 1985; Barkley, 1988; Trock, 1999; Bain and Golden, 2000; Schilling and Floden, 2015). While this is promising, further studies are needed to investigate whether the new probe set is more robust to taxon sampling.

The resulting tree topologies between Comp-1061 and Comp-ParaLoss-1272 were moderately incongruent (RF_{adj} = 0.625; Figure 4), indicating that species relationships varied

dependent on the probe set used. We suggest that these differences can be explained by 1) the different gene sets used to make the phylogeny, 2) the differences in paralog retention, or 3) the underlying biological processes present within *Packera*. First, given that this new probe set was complemented against Comp-1061 during production, there is no overlap of gene sequences between probe sets so only unique gene sequences, which have their own evolutionary histories, were used to generate each phylogeny. Therefore, the tree topologies and species relationships could differ since the Comp-ParaLoss-1272 phylogeny may be reflecting unique gene histories not shared with Comp-1061, and vice versa. Next, having fewer paralogs, as is seen in Comp-ParaLoss-1272, resulted in species relationships that may better reflect the underlying evolutionary histories and not as much gene heterogeneity (Smith and Hahn, 2021; Zhou et al., 2021). Finally, biological processes, such as hybridization, reticulation, or incomplete-lineage sorting (ILS), may be influencing our results as these processes are known to cause complications in phylogenetic construction (Arnold, 1997; Maddison, 1997; Alberts et al., 2002; Nussbaum et al., 2007).

Although only marginal, the Comp-ParaLoss-1272 tree had lower levels of discordance, indicating that Comp-ParaLoss-1272 provides more concordant nodes than Comp-1061, though the nodes are still highly discordant (Figures 5 and 6). It is reasonable to consider that the underlying biological processes discussed above may be influencing the level of discordance in our phylogeny, as *Packera* members have a long history of reticulation (e.g., Bremer, 1994; Bain et al., 1997) and hybridizing in the wild (e.g., Fernald, 1943; Barkley, 1962; Chapman et al., 1971; Uttal, 1984; Bain, 1988; Trock, 1999; Gramling, 2006; Weakley et al., 2011). Similar conclusions have been found in other groups (e.g., Sessa et al., 2012; Vargas et al., 2017; Morales-Briones et al., 2018). Interestingly, a recent study in *Packera* showed that low support

or discordant clades may be the result of ancient reticulation events in *Packera*'s history (Moore-Pollard and Mandel, 2023b), ultimately influencing the relationships and support within the species trees. We hypothesize that using Comp-ParaLoss-1272 will not only directly reduce issues associated with polyploidy, but also reduce issues from hybridization even if not addressed directly. Another possible explanation for the low node resolution is that only a subset of taxa (16 out of 88 *Packera* taxa) were used to generate these phylogenies. Having such low species sampling could influence species relationships and node support values given a lack of data (Heath et al., 2008; Sanderson et al., 2010).

Combining the sequence data from Comp-1061 with Comp-ParaLoss-1272, Comp-1061 + Comp-ParaLoss-1272, resulted in a topology that differed more substantially from the phylogeny generated using the Comp-1061 probe set ($RF_{adj} = 0.625$) compared to the Comp-ParaLoss-1272 probe set ($RF_{adj} = 0$) (Appendix S4). Additionally, Comp-1061 + Comp-ParaLoss-1272 resulted in a more resolved phylogeny than using Comp-1061 and Comp-ParaLoss-1272 alone (Appendix S4). For example, only three nodes had low support in the Comp-1061 + Comp-ParaLoss-1272 tree compared to four nodes in the Comp-ParaLoss-1272 only tree, and eight in the Comp-1061 only tree (Appendix S4). Even so, one of the discordant nodes in the combined tree had the lowest reported LPP value (LPP = 0.19), potentially indicating that underlying biological processes, such as hybridization or polyploidy, may be complicating the relationships at that node.

Ultimately, the most notable difference between the Comp-ParaLoss-1272 and Comp-1061 probe sets is the number of paralogs retained per individual, which was far fewer in the Comp-ParaLoss-1272 probe set than the Comp-1061. We predict this difference may be from 1) performing stricter filtering in the probe design process, 2) using more data to generate the probe

set, e.g., Comp-1061 used ESTs that were designed using low-coverage transcriptomes vs. Comp-ParaLoss-1272 which used complete transcriptomes, and 3) using more sequences across the phylogenetic breadth of the family, e.g., a single-copy gene in one lineage may be a multi-copy gene in a different lineage; therefore, using limited sampling when generating the Comp-1061 probe set (only three taxa in probe design) very likely missed some duplications that Comp-ParaLoss-1272 (48 taxa in probe design) was able to detect. While removing paralogs from a dataset may alleviate issues associated with ortholog determination in phylogenomic studies, it is important to note that paralogs are still reflective of the true evolutionary history of genes within some groups, including *Packera*. For example, hybridization and polyploidy are common in *Packera*, with around 40% of all *Packera* members exhibiting polyploidy (Trock, 1999, Moore-Pollard and Mandel, 2023a; Moore-Pollard and Mandel, 2023b), and as such paralogs are expected in the dataset as it reflects the true evolutionary history of the group. Therefore, removing paralogs can remove full gene histories, impacting your ability to accurately model processes like reticulation and polyploidy. Combining sequence data from both Comp-1061 and Comp-ParaLoss-1272 may be ideal if investigating clades for signal of reticulation or gene and genome duplications events. Additionally, new methods have been developed to better address these processes (Jackson et al., 2023; Morales-Briones et al., 2021; Nauheimer et al., 2021; Yang and Smith, 2014; Zhang and Mirarab, 2022), so we anticipate our combined probe set data will be useful for researchers who are interested in exploring their data in new ways. Even so, the Comp-1061 and Comp-ParaLoss-1272 probe sets are still comparable options for target-enrichment sequencing in lower-taxonomic members of Compositae.

Overall, the low paralog retention of the Comp-ParaLoss-1272 probe set can be very advantageous when dealing with groups known to be complicated by polyploidy since

polyploidy is typically associated with higher paralog retention (Lynch and Conery, 2000; Wolfe, 2001; Veitia, 2005). More attention is being placed on polyploidy in non-model plant groups (e.g., Lim et al., 2008; Bellinger et al., 2022; Fernández et al., 2022), and the underlying challenges associated with it are becoming more well known (see Rothfels, 2021). Being able to address these challenges early in the phylogenomic pipeline can improve phylogenetic reconstructions and provide more confidence in data interpretations. Given this, we anticipate that future work will test this probe set across different taxonomic levels, given that this study only tested it at the generic level, and provide additional support for the utility of this probe set in complex groups in the sunflower family. We hope this design approach will be seen as a model for other complex systems.

ACKNOWLEDGEMENTS

The authors thank Matthew D. Pollard for his bioinformatic help. We also thank Brian Brunelle at Arbor Biosciences for his assistance and expertise with probe design. Additionally, we thank the University of Memphis High-Performance Cluster (HPC) administrators, Eric Spangler and Kristian Skjervold, for their assistance with the HPC and overall willingness to provide support. Finally, we thank Jane Grimwood at HudsonAlpha.

AUTHOR CONTRIBUTIONS

E.R.M.P. designed the probe set, generated and analyzed data, and wrote the manuscript. J.R.M. helped design the probe set. D.S.J. provided transcriptome data for probe design and funds for sequencing. J.R.M. and D.S.J. provided edits to the manuscript.

All authors approved of the final version.

DATA AVAILABILITY STATEMENT

Raw sequence data are available in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (Bioprojects: PRJNA978591, PRJNA907383, and PRJNA994483).

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

Appendix S1. Voucher specimens to develop the probe set using MarkerMiner. Species names and authorities assigned by IPNI.

Appendix S2. List of 1,925 targeted loci in the Compositae-ParaLoss-1272 probe set and information about their associated functions in *Arabidopsis thaliana* L. (source: The *Arabidopsis*

Information Resource (TAIR); https://www.arabidopsis.org/tools/bulk/genes/index.jsp). *Vitis vinifera* L. specific genes that have no known function (n = 17) are included.

Appendix S3. Table of HybPiper summary statistics for the six Asteraceae genomes from the CapSim run.

Appendix S4. Tanglegrams comparing the relationships between the combined dataset, Compositae-1061 + Compositae-ParaLoss-1272, against the individual datasets: Compositae-1061 (**A**) and Compositae-ParaLoss-1272 (**B**). Lines between the taxa at the tips compare relationships: solid line indicates the same relationship; dashed line indicates differing relationships. Local posterior probability (LPP) values are represented at each node and colored accordingly: full support (1,0LPP) is blue, moderate support (0.9-0.99LPP) is green, while low support (≤0.89LPP) is red.

Appendix S5. Tanglegrams comparing the relationships between a pruned down version of the Moore-Pollard and Mandel (2023a) tree now containing the 19 taxa used in this study, compared to the Compositae-1061 (**A**) and Compositae-ParaLoss-1272 (**B**) trees generated in this study. Lines between the taxa at the tips compare relationships: solid line indicates the same relationship; dashed line indicates differing relationships.

Appendix S6. General and full HybPiper stats of the Illumina sequence run.

Appendix S7. Compositae-ParaLoss-1272 probe set file for bioinformatic analyses.

514 **REFERENCES**

515 Acha, S., and L. C. Majure. 2022. A new approach using targeted sequence capture for
516 phylogenomic studies across Cactaceae. *Genes* 13: 350.
517 Alberts, B., A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walte. 2002. Molecular Biology of
518 the Cell. 4th Ed. New York: Garland Science.
519 Arnold, M. L. 1997. Natural hybridization and evolution, Oxford Series in Ecology and
520 Evolution. Oxford University Press, New York, NY.
521 Bain, J. F. 1988. Taxonomy of *Senecio streptanthifolius* Greene. *Rhodora* 90: 277–312.
522 Bain, J. F., and R. K. Jansen. 1995. A phylogenetic analysis of the aureoid *Senecio* (Asteraceae)
523 complex based on ITS sequence data. *Plant Systematics and Evolution* 195: 209–219.
524 Bain, J. F., B. S. Tyson, and D. F. Bray. 1997. Variation in pollen wall ultrastructure in New
525 World Senecioneae (Asteraceae), with special reference to *Packera*. *Canadian Journal of*
526 *Botany* 75: 730–735.
527 Bain, J. F., and J. L. Golden. 2000. A phylogeny of *Packera* (Senecioneae; Asteraceae) based on
528 internal transcribed spacer region sequence data and a broad sampling of outgroups.
529 *Molecular Phylogenetics and Evolution* 16: 331–338.
530 Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, et
531 al. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell
532 sequencing. *Journal of Computational Biology* 19: 455–477.
533 Barkley, T. M. 1962. A Revision of *Senecio aureus*. *Transactions of the Kansas Academy of*
534 *Science* 65: 318–364.
535 Barkley, T. M. 1985. Infrageneric groups in *Senecio*, S. L., and *Cacalia*, S. L. (Asteraceae:
536 Senecioneae) in Mexico and Central America. *Brittonia* 37: 211–218.

537 Barkley, T. M. 1988. Variation among the Aureoid *Senecios* of North America: A geohistorical
538 interpretation. *Botanical Review* 54: 82–106.
539 Bellinger, M. R., E. M. Datlof, K. E. Selph, T. J. Gallaher, and M.L. Knope. 2022. A genome for
540 *Bidens hawaiensis*: A member of a hexaploid Hawaiian plant adaptive radiation. *Journal*
541 *of Heredity* 113: 205–214.
542 Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: A flexible trimmer for Illumina
543 sequence data. *Bioinformatics* 30: 2114–2120.
544 Badouin, H., J. Gouzy, C. J. Grassa, F. Murat, S. E. Staton, L. Cottret, C. Lelandais-Brière, et al.
545 2017. The Sunflower genome provides insights into oil metabolism, flowering and
546 Asterid evolution. *Nature* 546: 148–52.
547 Bremer, K. 1994. Tribe Senecioneae. In Asteraceae: Cladistics and Classification, 479–520.
548 Timber Press, Portland.
549 Brown, J. W., J. F. Walker, and S. A. Smith. 2017. Phyx: Phylogenetic tools for unix.
550 *Bioinformatics* 33: 1886–1888.
551 Cao, M. D., D. Ganesamoorthy, C. Zhou, and L. J. M. Coin. 2018. Simulating the dynamics of
552 targeted capture sequencing with CapSim. *Bioinformatics* 34: 873–874.
553 Chamala, S., N. García, G. T. Godden, V. Krishnakumar, I. E. Jordon-Thaden, R. De Smet, W.
554 B. Barbazuk, et al. 2015. MarkerMiner 1.0: A new application for phylogenetic marker
555 development using angiosperm transcriptomes. *Applications in Plant Sciences* 3:
556 1400115.
557 Chapman, G. C., and S. B. Jones. 1971. Hybridization between *Senecio smallii* and *S.*
558 *tomentosus* (Compositae) on the granitic flatrocks of the Southeastern United States.
559 *Brittonia* 23: 209–216.

560 Chapman, M. A. 2015. Transcriptome sequencing and marker development for four
561 underutilized Legumes. *Applications in Plant Sciences* 3: 1400111.
562 Danecek, P., J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, et al.
563 2021. Twelve years of SAMtools and BCFtools. *Gigascience* 10: giab008.
564 de Lima Ferreira, P., R. Batista, T. Andermann, M. Groppo, C. D. Bacon, and A. Antonelli.
565 2022. Target sequence capture of Barnadesioideae (Compositae) demonstrates the utility
566 of low coverage loci in phylogenomic analyses. *Molecular Phylogenetics and Evolution*
567 169: 107432.
568 De Smet, R., K. L. Adams, K. Vandepoele, M. C. E. Van Montagu, S. Maere, and Y. Van De
569 Peer. 2013. Convergent gene loss following gene and genome duplications creates single-
570 copy families in flowering plants. *Proceedings of the National Academy of Sciences of*
571 *the United States* 110: 2898–2903.
572 Duncavage, E. J., J. F. Coleman, M. E. de Baca, S. Kadri, A. Leon, M. Routbort, S. Roy, et al.
573 2023. Recommendations for the use of in silico approaches for next-generation
574 sequencing bioinformatic pipeline validation: A joint report of the association for
575 molecular pathology, association for pathology informatics, and college of American
576 pathologists. *Journal of Molecular Diagnostics* 25: 3–16.
577 Emms, D. M., and S. Kelly. 2015. OrthoFinder: solving fundamental biases in whole genome
578 comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* 16:
579 1–14.
580 Emms, D. M., and S. Kelly. 2019. OrthoFinder: Phylogenetic orthology inference for
581 comparative genomics. *Genome Biology* 20: 1–14.

582 Eserman, L. A., S. K. Thomas, E. E. D. Coffey, and J. H. Leebens-Mack. 2021. Target sequence
583 capture in orchids: Developing a kit to sequence hundreds of single-copy loci.
584 *Applications in Plant Science* 9: 1–8.
585 Fernald, M. L. 1943. Virginia botanizing under restrictions. *Rhodora* 45: 485–511.
586 Fernández, P., O. Hidalgo, A. Juan, I. J. Leitch, A. R. Leitch, L. Palazzesi, L. Pegoraro, J. Viruel,
587 and J. Pellicer. 2022. Genome insights into autopolyploid evolution: A case study in
588 *Senecio doronicum* (Asteraceae) from the southern alps. *Plants* 11: 1235.
589 Folk, R. A., J. R. Mandel, and J. V. Freudenstein. 2015. A protocol for targeted enrichment of
590 intron-containing sequence markers for recent radiations: A phylogenomic example from
591 *Heuchera* (Saxifragaceae). *Applications in Plant Sciences* 3: 1–10.
592 Fonseca, L. H. M., and L. G. Lohmann. 2020. Exploring the potential of nuclear and
593 mitochondrial sequencing data generated through genome-skimming for plant
594 phylogenetics: A case study from a clade of neotropical lianas. *Journal of Systematics*
595 *and Evolution* 58: 18–32.
596 Freeman, C. C. 1985. A revision of the aureiod species of *Senecio* (Asteraceae: Senecioneae) in
597 Mexico, with a cytogeographic and phylogenetic interpretation of the aureoid complex.
598 Dissertation, Kansas State University, USA.
599 Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, et
600 al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference
601 genome. *Nature Biotechnology* 29: 644–52.
602 Gramling, A. 2006. A conservation assessment of *Packera millefolium*, a Southern Appalachian
603 Endemic. Dissertation, University of North Carolina at Chapel Hill, USA.
604

605 Haas, B. J., A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger,
606 et al. 2013. De novo transcript sequence reconstruction from RNA-Seq using the Trinity
607 platform for reference generation and analysis. *Nature Protocols* 8: 1494–1512.
608 Heath, T. A., S. M. Hedtke, and D. M. Hillis. 2008. Taxon sampling and the accuracy of
609 phylogenetic analyses. *Journal of Systematics and Evolution* 48: 239–257.
610 Huang, C. H., C. Zhang, M. Liu, Y. Hu, T. Gao, J. Qi, and H. Ma. 2016. Multiple
611 polyploidization events across Asteraceae with two nested events in the early history
612 revealed by nuclear phylogenomics. *Molecular Biology and Evolution* 33: 2820–2835.
613 Jackson, C., T. McLay, and A. N. Schmidt-Lebuhn. 2023. hybpiper-nf and paragone-nf:
614 Containerization and additional options for target capture assembly and paralog
615 resolution. *Applications in Plant Sciences* 11: e11532.
616 Johnson, M. G., E. M. Gardner, Y. Liu, R. Medina, B. Goffinet, A. J. Shaw, N. J. C. Zerega, and
617 N. J. Wickett. 2016. HybPiper: Extracting coding sequence and introns for phylogenetics
618 from high-throughput sequencing reads using target enrichment. *Applications in Plant*
619 *Sciences* 4: 1600016.
620 Johnson, M.G., L. Pokorny, S. Dodsworth, L. R. Botigué, R. S. Cowan, A. Devault, W. L.
621 Eiserhardt, et al. 2019. A universal probe set for targeted sequencing of 353 nuclear
622 genes from any flowering plant designed using k-medoids clustering. *Systematic Biology*
623 68: 594–606.
624 Katoh, K., and D. M. Standley. 2013. MAFFT multiple sequence alignment software version 7:
625 Improvements in performance and usability. *Molecular Biology and Evolution* 30: 772–
626 780.

627 Kück, P., and G. C. Longo. 2014. FASconCAT-G: Extensive functions for multiple sequence
628 alignment preparations concerning phylogenetic studies. *Frontiers in Zoology* 11: 1–8.
629 Kumar, S., G. Stecher, M. Li, C. Knyaz, and K. Tamura. 2018. MEGA X: Molecular
630 evolutionary genetics analysis across computing platforms. *Molecular Biology and*
631 *Evolution* 35: 1547–1549.
632 Langmead, B., and S. L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. *Nature*
633 *Methods* 9: 357–359.
634 Langmead, B., C. Wilks, V. Antonescu, and R. Charles. 2019. Scaling read aligners to hundreds
635 of threads on general-purpose processors. *Bioinformatics* 35: 421–432.
636 Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler
637 transform. *Bioinformatics* 25: 1754–1760.
638 Lichter-Marck, I. H., W. A. Freyman, C. M. Siniscalchi, J. R. Mandel, A. Castro-Castro, G.
639 Johnson, and B. G. Baldwin. 2020. Phylogenomics of *Perityleae* (Compositae) provides
640 new insights into morphological and chromosomal evolution of the rock daisies. *Journal*
641 *of Systematics and Evolution* 58: 853–880.
642 Lim, K. Y., D. E. Soltis, P. S. Soltis, J. Tate, R. Matyasek, H. Srubarova, A. Kovarik, et al. 2008.
643 Rapid chromosome evolution in recently formed polyploids in *Tragopogon* (Asteraceae).
644 *PLoS One* 3: e3353.
645 Liu, Y., M. G. Johnson, C. J. Cox, R. Medina, N. Devos, A. Vanderpoorten, L. Hedenäs, et al.
646 2019. Resolution of the ordinal phylogeny of mosses using targeted exons from
647 organellar and nuclear genomes. *Nature Communications* 10: 1485.
648 Lynch, M., and J. S. Conery. 2000. The evolutionary fate and consequences of duplicate genes.
649 *Science* 290: 1151–1155.

650 Maddison, W. P. 1997. Gene trees in species trees. *Systematic Biology* 46: 523–536.
651 Mandel, J. R., R. B. Dikow, V. A. Funk, R. R. Masalia, S. E. Staton, A. Kozik, R. W.
652 Michelmore, et al. 2014. A target enrichment method for gathering phylogenetic
653 information from hundreds of loci: An example from the Compositae. *Applications in*
654 *Plant Science* 2: 1300085.
655 Mandel, J. R., M. S. Barker, R. J. Bayer, R. B. Dikow, T. G. Gao, K. E. Jones, S. Keeley, et al.
656 2017. The Compositae Tree of Life in the age of phylogenomics. *Journal of Systematics*
657 *and Evolution* 55: 405–410.
658 Mandel, J. R., R. B. Dikow, C. M. Siniscalchi, R. Thapa, L. E. Watson, and V.A. Funk. 2019. A
659 fully resolved backbone phylogeny reveals numerous dispersals and explosive
660 diversifications throughout the history of Asteraceae. *Proceedings of the National*
661 *Academy of Sciences of the United States* 116: 14083–14088.
662 May, V., L. Koch, B. Fischer-Zirnsak, D. Horn, P. Gehle, U. Kornak, D. Beule, and M.
663 Holtgrewe. 2022. ClearCNV: CNV calling from NGS panel data in the presence of
664 ambiguity and noise. *Bioinformatics* 38: 3871–3876.
665 McKain, M. R., M. G. Johnson, S. Uribe-Convers, D. Eaton, and Y. Yang. 2018. Practical
666 considerations for plant phylogenomics. *Applications in Plant Sciences* 6: 1–15.
667 Moore-Pollard, E. R., and J. R. Mandel. 2023a. Resolving evolutionary relationships in the
668 groundsels: phylogenomics, divergence time estimates, and biogeography of *Packera*
669 (Asteraceae: Senecioneae). *bioRxiv*. DOI: 10.1101/2023.07.18.549592.
670 Moore-Pollard, E. R., and J. R. Mandel. 2023b. From paralogy to hybridization: Investigating
671 causes of underlying phylogenomic discordance using the complex
672 genus *Packera* (Senecioneae; Asteraceae). *bioRxiv*. DOI: 10.1101/2023.08.14.553290.

673 Morales-Briones, D. F., A. Liston, and D. C. Tank. 2018. Phylogenomic analyses reveal a deep
674 history of hybridization and polyploidy in the Neotropical genus *Lachemilla* (Rosaceae).
675 *New Phytology* 218: 1668–1684.
676 Morales-Briones, D. F., G. Kadereit, D. T. Tefarikis, M. J. Moore, S. A. Smith, S. F.
677 Brockington, A. Timoneda, et al. 2021. Disentangling sources of gene tree discordance in
678 phylogenomic data sets: Testing ancient hybridizations in Amaranthaceae s.L. *Systematic*
679 *Biology* 70: 219–235.
680 Müller, R., A. Bleckmann, and R. Simon. 2008. The receptor kinase *CORYNE* of *Arabidopsis*
681 transmits the stem cell-limiting signal *CLAVATA3* independently of *CLAVATA1*. *The*
682 *Plant Cell* 20: 934–46.
683 Nauheimer, L., N. Weigner, E. Joyce, D. Crayn, C. Clarke, and K. Nargar. 2021. HybPhaser: a
684 workflow for the detection and phasing of hybrids in target capture datasets. *Applications*
685 *in Plant Sciences* 9: 1–14.
686 Nussbaum, S., R. R. McInnes, and H. F. Willard. 2007. Genetics in Medicine. Saunders
687 Elselvier, Philadelphia, PA, USA.
688 One Thousand Plant Transcriptomes Initiative. 2019. One thousand plant transcriptomes and the
689 phylogenomics of green plants. *Nature* 574: 679–685.
690 Palazzesi, L., J. Pellicer, V. D. Barreda, B. Loeuille, J. R. Mandel, L. Pokorny, C. M. Siniscalchi,
691 et al. 2022. Asteraceae as a model system for evolutionary studies: from fossils to
692 genomes. *Botanical Journal of the Linnean Society* 200: 143–164.
693 Pease, J. B., J. W. Brown, J. F. Walker, C. E. Hinchliff, and S. A. Smith. 2018. Quartet Sampling
694 distinguishes lack of support from conflicting support in the green plant tree of life.
695 *American Journal of Botany* 105: 385–403.

696 Pelser, P. B., B. Nordenstam, J. W. Kadereit, and L. E. Watson. 2007. An ITS phylogeny of tribe
697 Senecioneae (Asteraceae) and a new delimitation of *Senecio* L. *Taxon* 56: 1077–1104.
698 R Core Team. 2016. R: A language and environment for statistical computing. R Foundation for
699 Statistical Computing, Vienna, R Foundation for Statistical Computing, Vienna.
700 Reichelt, N., J. Wen, C. Pätzold, and M. S. Appelhans. 2021. Target enrichment improves
701 phylogenetic resolution in the genus *Zanthoxylum* (Rutaceae) and indicates both
702 incomplete lineage sorting and hybridization events. *Annals of Botany* 128: 497–510.
703 Revell, L. J. 2012. phytools: An R package for phylogenetic comparative biology (and other
704 things). *Methods in Ecology and Evolution* 3: 217–223.
705 Rothfels, C. J. 2021. Polyploid phylogenetics. *New Phytologist* 230: 66–72.
706 RStudio. 2020. RStudio: Integrated Development for R. RStudio, PBC, Boston, MA, RStudio,
707 PBC, Boston, MA.
708 Sanderson, M. J., M. M. McMahon, and M. Steel. 2010. Phylogenomics with incomplete taxon
709 coverage: the limits to inference. *BMC Evolutionary Biology* 10: 155.
710 Sayyari, E., and S. Mirarab. 2016. Fast coalescent-based computation of local branch support
711 from quartet frequencies. *Molecular Biology and Evolution* 33: 1654–1668.
712 Schilling, E. E., and A. Floden. 2015. Barcoding the Asteraceae of Tennessee, tribe Cichorieae.
713 *Phytoneuron* 19: 1–8.
714 Schliep, K. P. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics* 27: 592–593.
715 Sessa, E. B., E. A. Zimmer, and T. J. Givnish. 2012. Reticulate evolution on a global scale: a
716 nuclear phylogeny for New World Dryopteris (Dryopteridaceae). *Molecular*
717 *Phylogenetics and Evolution* 64: 563–581.

718 Shah, T., J. V. Schneider, G. Zizka, O. Maurin, W. Baker, F. Forest, G. E. Brewer, et al. 2021.
719 Joining forces in Ochnaceae phylogenomics: a tale of two targeted sequencing probe kits.
720 *American Journal of Botany* 108: 1201–1216.
721 Siniscalchi, C. M., B. Loeuille, V. A. Funk, J. R. Mandel, and J. R. Pirani. 2019. Phylogenomics
722 yields new insight into relationships within Vernonieae (Asteraceae). *Frontiers in Plant*
723 *Science* 10: 1–16.
724 Siniscalchi, C. M., O. Hidalgo, L. Palazzesi, J. Pellicer, L. Pokorny, O. Maurin, I. J. Leitch, et al.
725 2021. Lineage-specific vs. universal: A comparison of the Compositae1061 and
726 Angiosperms353 enrichment panels in the sunflower family. *Applications in Plant*
727 *Sciences* 9.
728 Siniscalchi, C. M., J. Ackerfield, and R. A. Folk. 2023. Diversification and biogeography of
729 North American thistles (*Cirsium*: Carduoideae: Compositae): drivers of a rapid
730 continent-wide radiation. *International Journal of Plant Sciences* 184.
731 Smith, M. L., and M. W. Hahn. 2021. New approaches for inferring phylogenies in the presence
732 of paralogs. *Trends in Genetics* 37: 174–187.
733 Smith, S. A., M. J. Moore, J. W. Brown, and Y. Yang. 2015. Analysis of phylogenomic datasets
734 reveals conflict, concordance, and gene duplications with examples from animals and
735 plants. *BMC Evolutionary Biology* 15: 1–15.
736 Stamatakis, A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of
737 large phylogenies. *Bioinformatics* 30: 1312–1313.
738 Straub, S. C. K., M. Parks, K. Weitemier, M. Fishbein, R. C. Cronn, and A. Liston. 2012.
739 Navigating the tip of the genomic iceberg: Next-generation sequencing for plant
740 systematics. *American Journal of Botany* 99: 349–364.

741 Stull, G. W., M. J. Moore, V. S. Mandala, N. A. Douglas, H. Kates, X. Qi, S. F. Brockington, et
742 al. 2013. A targeted enrichment strategy for massively parallel sequencing of angiosperm
743 plastid genomes. *Applications in Plant Sciences* 1: 1200497.
744 Thapa, R., R. J. Bayer, and J. R. Mandel. 2020. Phylogenomics resolves the relationships within
745 *Antennaria* (Asteraceae, Gnaphalieae) and yields new insights into its morphological
746 character evolution and biogeography. *Systematic Botany* 45: 387–402.
747 Trock, D. K. 1999. A revisionary synthesis of the genus *Packera* (Asteraceae: Senecioneae).
748 Dissertation, Kansas State University, USA.
749 Uttal, L. J. 1984. *Senecio millefolium* T. & G. (Asteraceae) and its introgressants. *SIDA*
750 *Contributions to Botany* 10: 216–222.
751 Vargas, O. M., E. M. Ortiz, and B. B. Simpson. 2017. Conflicting phylogenomic signals reveal a
752 pattern of reticulate evolution in a recent high-Andean diversification (Asteraceae:
753 Astereae: *Diplostephium*). *New Phytologist* 214: 1736–1750.
754 Vatanparast, M., A. Powell, J. J. Doyle, and A. N. Egan. 2018. Targeting legume loci: A
755 comparison of three methods for target enrichment bait design in Leguminosae
756 phylogenomics. *Applications in Plant Sciences* 6: e1036.
757 Veitia, R. A. 2005. Paralogs in polyploids: One for all and all for one? *Plant Cell* 17: 4–11.
758 Villaverde, T., L. Pokorny, S. Olsson, M. Rincón-Barrado, M. G. Johnson, E. M. Gardner, N. J.
759 Wickett, et al. 2018. Bridging the micro- and macroevolutionary levels in
760 phylogenomics: Hyb-Seq solves relationships from populations to species and above.
761 *New Phytologist* 220: 636–650.
762 Weakley, A. S., R. J. LeBlond, B. A. Sorrie, C. T. Witsell, L. D. Estes, K. Gandhi, K. G.
763 Mathews, and A. Ebihara. 2011. New combinations, rank changes, and nomenclatural

764 and taxonomic comments in the vascular flora of The Southeastern United States.
765 *Journal of the Botanical Research Institute of Texas* 5: 437–455.
766 Weigel, D., J. Alvarez, D. R. Smyth, M. F. Yanofsky, and E. M. Meyerowitz. 1992. *LEAFY*
767 controls floral meristem identity in *Arabidopsis*. *Cell* 69: 843–59.
768 Weitemier, K., S. C. K. Straub, R. C. Cronn, M. Fishbein, R. Schmickl, A. McDonnell, and A.
769 Liston. 2014. Hyb-Seq: Combining target enrichment and genome skimming for plant
770 phylogenomics. *Applications in Plant Sciences* 2: 1400042.
771 Wickham, H. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York,
772 Springer-Verlag New York.
773 Wolf, P. G., T. A. Robison, M. G. Johnson, M. A. Sundue, W. L. Testo, and C. J. Rothfels. 2018.
774 Target sequence capture of nuclear-encoded genes for phylogenetic analysis in ferns.
775 *Applications in Plant Sciences* 6: 1–8.
776 Wolfe, K. H. 2001. Yesterday's polyploids and the mystery of diploidization. *Nature Reviews*
777 *Genetics* 2: 333–341.
778 Yang, Y., and S. A. Smith. 2014. Orthology inference in nonmodel organisms using
779 transcriptomes and low-coverage genomes: Improving accuracy and matrix occupancy
780 for phylogenomics. *Molecular Biology and Evolution* 31: 3081–3092.
781 Zhang, C., M. Rabiee, E. Sayyari, and S. Mirarab. 2018. ASTRAL-III: Polynomial time species
782 tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* 19: 15–30.
783 Zhang, C., C.-H. Huang, M. Liu, Y. Hu, J. L. Panero, F. Luebert, T. Gao, and H. Ma. 2021.
784 Phylotranscriptomic insight into Asteraceae diversity, polyploidy, and morphological
785 innovation. *Journal of Integrative Plant Biology* 63: 1273–1293.

786 Zhang, C., and S. Mirarab. 2022. ASTRAL-Pro 2: ultrafast species tree reconstruction from
787 multi-copy gene family trees. *Bioinformatics* 38: 4949–4950.
788 Zhou, W., J. Soghigian, and Q. Y. Xiang. 2021. A new pipeline for removing paralogs in target
789 enrichment data. *Systematic Biology* 71: 410-425.
790

Table 1. Table listing the major differences between the sunflower-family specific probe sets, Compositae-ParaLoss-1272 (Comp-ParaLoss-1272) and Compositae-1061 (Comp-1061), and the angiosperm-wide probe set, Angiosperm-353 (Angio-353). Rows can be defined as: # loci = number of targeted loci; # baits = number of baits in probe set; # loci overlap = number of loci that overlap with another probe set indicated within parentheses; # species = number of species used to develop probe set; Input data = input data type to develop probe set; Tool = tool use to develop probe set.

	Comp-ParaLoss-1272	**Comp-1061**	**Angio-353**
# loci	1,272	1,061	353
# baits	60,158	9,678†	75,151‡
# loci overlap	0	30 (with Angio-353) §	30 (with Comp-1061) §
# species	48	3	42
Input data	transcriptomes	Expressed sequence tags (EST)	transcriptomes
Tool	MarkerMiner	BLAST	k-medoid clustering

† Mandel, J. R., R. B. Dikow, V. A. Funk, R. R. Masalia, S. E. Staton, A. Kozik, R. W. Michelmore, et al. 2014. A target enrichment method for gathering phylogenetic information from hundreds of loci: An example from the Compositae. *Applications in Plant Science* 2: 1300085.

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

§ Siniscalchi, C. M., O. Hidalgo, L. Palazzesi, J. Pellicer, L. Pokorny, O. Maurin, I. J. Leitch, et al. 2021. Lineage-specific vs. universal: A comparison of the Compositae1061 and Angiosperms353 enrichment panels in the sunflower family. *Applications in Plant Sciences* 9.

Table 2. Voucher specimens for the Illumina sequence run. Publication status and authorities assigned by IPNI. * indicates a report for only the Compositae-ParaLoss-1272 probe set.

Species	Location	Collector and # (Herbarium)	Coll. date	Sheet barcode or ID number	Raw reads (paired)*	Reads mapped*	NCBI accession - Comp-1061	NCBI accession - Comp-ParaLoss-1272
Emilia fosbergii Nicolson	USA; FL, Osceola County	Wayne D. Longbottom, David H. Williams, Holly L. Williams 14545 (NY)	18-Nov-10	02074297	1,572,629,062	10,414,762	SRR22543392	SRR24860889
Packera aurea (L.) Á.Löve & D.Löve	USA; Tennessee, Campbell County	Floden 866 (TENN)	s.d.	N/A	3,009,238,834	19,928,734	SRR22543326	SRR24860888
Packera cana (Hook.) W.A.Weber & Á.Löve	USA; ID, Adams County	Don Knoke 2101 (WTU)	25-Jun-11	406472	4,989,136,942	33,040,642	SRR24862023	SRR24860878
Packera candidissima (Greene) W.A.Weber & Á.Löve	Mexico; Sierra Madre Occidental, Mexico	Robert A. Bye 9680 (ASU)	26-May-80	121438	2,880,272,150	19,074,650	SRR22543387	SRR24860877
Packera castoreus (S.L.Welsh) Kartesz	USA; UT, Piute County	Alan Taye 3674 (OSC)	20-Sep-87	172202	2,269,567,448	15,030,248	SRR22543385	SRR24860876
Packera crocata (Rydb.) W.A.Weber & Á.Löve	USA; CO, Jackson County	Mary Damm 38 (OSC)	29-Jul-02	244322	6,132,282,442	40,611,142	SRR22543379	SRR24860875
Packera cynthioides (Greene) W.A.Weber & Á.Löve	USA; NM, Grant County	Darrell E. Ward 80-010 (NY)	6-Sep-80	03088483	2,917,414,224	19,320,624	SRR22543377	SRR24860874
Packera dubia (Spreng.) Trock & Mabb.	USA; NC, Chesapeake County	J. Brandon Fuller (NCU)	29-Jun-20	N/A	2,167,035,730	14,351,230	SRR22543313	SRR24860880
Packera franciscana (Greene) W.A.Weber & Á.Löve	USA; AZ, Coconino County	J. Resinger 1577 (ARIZ)	14-Jul-76	233800	4,604,239,452	30,491,652	SRR22543368	SRR24860873
Packera glabella (Poir.) C.Jeffrey	USA; Tennessee, Bradley County	DeSelm 06-04 (TENN)	s.d.	N/A	3,641,082,026	24,113,126	SRR22543366	SRR24860872
Packera greenei (A.Gray) W.A.Weber & Á.Löve	USA; CA, Trinity County	E.R. Moore 8 (MEM)	27-Jun-19	20904	2,943,301,060	19,492,060	SRR22543365	SRR24860871
Packera layneae (Greene) W.A.Weber & Á.Löve	USA; CA, El Dorado County	Kathryn A. Beck 200310 (WTU)	30-Apr-03	375035	5,681,052,766	37,622,866	SRR22543356	SRR24860887
Packera loratifolia (Greenm.) W.A.Weber & Á.Löve	Mexico; Sierra La Viga, Mexico	J.A. Villarreal, J. Valdes R 5163 (ASU)	16-Sep-89	182928	2,487,875,698	16,475,998	SRR22543355	SRR24860886
Packera musiniensis (S.L.Welsh) Trock	USA; UT, Sanpete County	D. Atwood 21259 (ARIZ)	9-Aug-96	334839	2,988,242,284	19,789,684	SRR22543346	SRR24860885
Packera porteri (Greene) C.Jeffrey	USA; OR, County	Coll. Wm. Cusick 2308 (OSC)	8/3/1899	97915	4,421,594,684	29,282,084	SRR22543334	SRR24860884
Packera pseudaurea (Rydb.) W.A.Weber & Á.Löve	USA; ID, Valley County	James F. Smith 9147 (OSC)	29-Jul-10	228940	3,922,950,102	25,979,802	SRR22543332	SRR24860883
Packera streptanthifolia (Greene) W.A.Weber & Á.Löve	USA; OR, Grant County	Sharon Birks 2010-42 (OSC)	16-Jul-10	255384	13,606,754,356	90,110,956	SRR22543319	SRR24860882

Packera texensis O'Kennon & Trock	USA; TX, Gillespie County	B.L. Turner 24-75 (TEX)	10-Apr-04	00211804	4,920,515,898	32,586,198	SRR22543316	SRR24860881
Roldana gilgii (Greenm.) H.Rob. & Brettell	Mexico; Chiapas, Mexico	D.E. Breedlove 24411 (TEX)	5-Mar-72	00062617	2,082,647,568	13,792,368	SRR22543307	SRR24860879

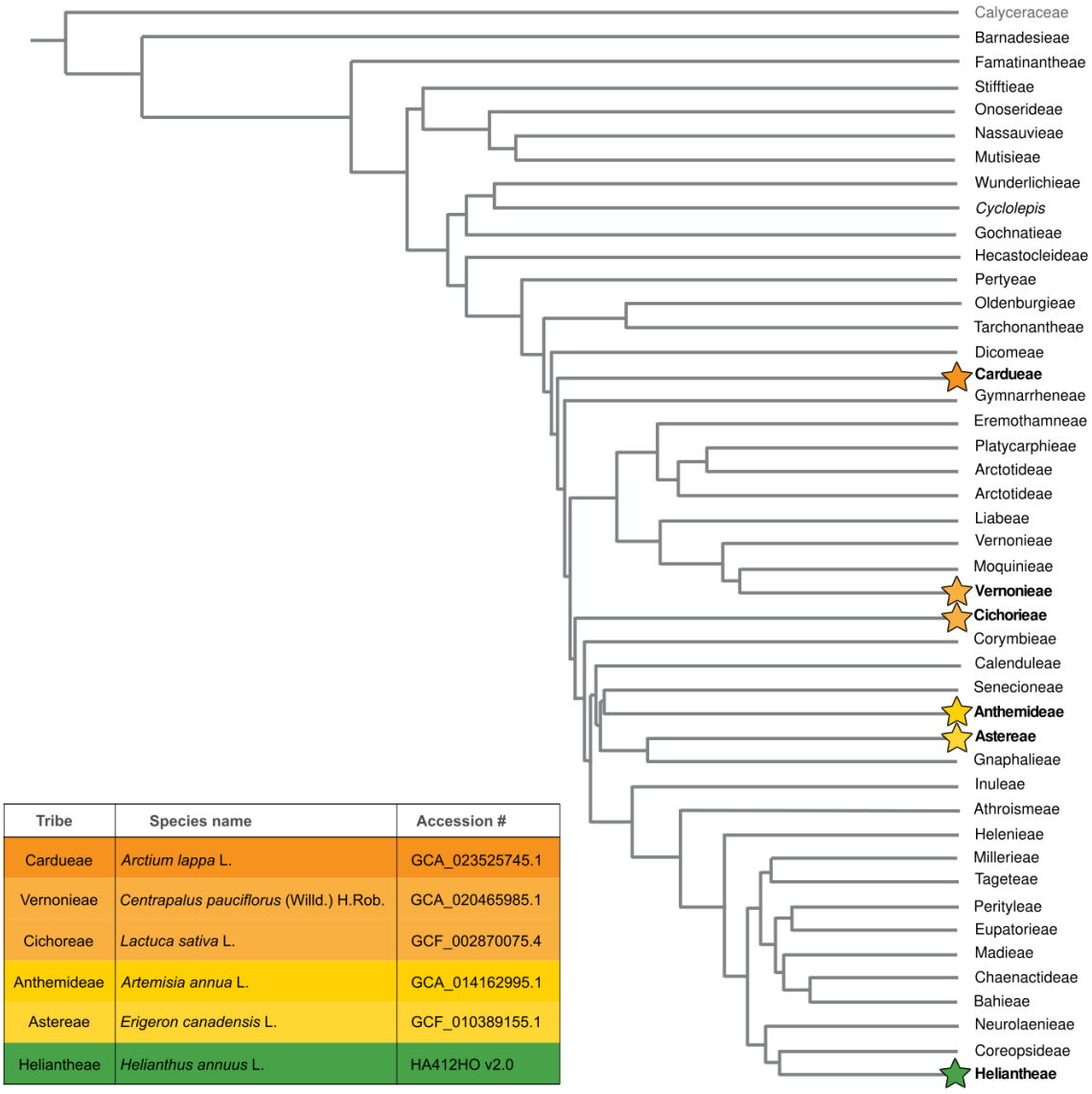
Table 3. Summary statistics of the CapSim run after running the 'stats' function in HybPiper.

Species	Comp-1061					Comp-ParaLoss-1272				
	Reads mapped	% on target	Genes mapped	% genes retained	Paralog warnings	Reads mapped	% on target	Genes mapped	% Genes retained	Paralog warnings
Artemisia annua L.	93,739,367	93.7%	407	38.4%	108	97,421,399	97.4%	433	34.0%	1
Helianthus annuus L.	97,351,903	97.4%	750	70.7%	250	97,357,378	97.4%	403	31.7%	1
Centrapalus pauciflorus (Willd.) H.Rob.	94,823,613	94.8%	466	43.9%	101	97,708,408	97.7%	468	36.8%	0
Lactuca sativa L.	98,218,579	98.2%	749	70.6%	223	97,532,753	97.5%	519	40.8%	2
Erigeron canadensis L.	95,987,418	96.0%	548	51.6%	96	97,231,893	97.2%	500	39.3%	1
Arctium lappa L.	92,956,716	93.0%	388	36.6%	103	97,530,647	97.5%	399	31.4%	1

Table 4. Summary statistics of the Illumina sequencing run after running the 'stats' function in HybPiper.

Species	Comp-1061					Comp-ParaLoss-1272				
	Reads mapped	% on target	Genes mapped	% genes retained	Paralog warnings	Reads mapped	% on target	Genes mapped	% genes retained	Paralog warnings
Emilia fosbergii Nicolson	1,185,704	59%	1006	94.8%	95	11,236,129	65%	1259	99.0%	14
Packera aurea (L.) Á.Löve & D.Löve	5,184,671	53%	1016	95.8%	214	4,438,388	26%	1265	99.4%	9
Packera cana (Hook.) W.A.Weber & Á.Löve	1,532,039	21%	997	94.0%	130	8,297,634	35%	1268	99.7%	7
Packera candidissima (Greene) W.A.Weber & Á.Löve	558,742	36%	999	94.2%	91	5,690,438	43%	1264	99.4%	5
Packera castoreus (S.L.Welsh) Kartesz	1,654,718	37%	1043	98.3%	347	2,912,785	32%	1260	99.1%	2
Packera crocata (Rydb.) W.A.Weber & Á.Löve	2,361,884	36%	1021	96.2%	265	10,762,526	35%	1267	99.6%	11
Packera cynthioides (Greene) W.A.Weber & Á.Löve	1,171,793	29%	1007	94.9%	150	2,064,556	36%	1258	98.9%	4
Packera dubia (Spreng.) Trock & Mabb.	4,514,739	39%	1016	95.8%	233	2,775,445	26%	1266	99.5%	5
Packera franciscana (Greene) W.A.Weber & Á.Löve	1,573,692	41%	992	93.5%	256	9,169,648	45%	1264	99.4%	7
Packera glabella (Poir.) C.Jeffrey	1,972,057	34%	1029	97.0%	256	6,012,371	31%	1266	99.5%	10
Packera greenei (A.Gray) W.A.Weber & Á.Löve	2,024,706	34%	1013	95.5%	250	4,102,840	27%	1259	99.0%	8
Packera layneae (Greene) W.A.Weber & Á.Löve	2,814,096	35%	1048	98.8%	394	8,240,509	26%	1268	99.7%	8
Packera loratifolia (Greenm.) W.A.Weber & Á.Löve	511,859	43%	1001	94.3%	53	2,435,806	35%	1262	99.2%	0
Packera musiniensis (S.L.Welsh) Trock	68,064	9%	923	87.0%	35	6,518,518	44%	1254	98.6%	9
Packera porteri (Greene) C.Jeffrey	1,510,836	39%	1018	95.9%	193	5,896,137	40%	1268	99.7%	6
Packera pseudaurea (Rydb.) W.A.Weber & Á.Löve	3,914,039	41%	1027	96.8%	309	7,423,481	41%	1264	99.4%	13

Packera streptanthifolia (Greene) W.A.Weber & Á.Löve	2,695,188	38%	1026	96.7%	309	23,885,890	39%	1271	99.9%	14
Packera texensis O'Kennon & Trock	2,516,755	33%	1008	95.0%	238	9,026,502	38%	1266	99.5%	12
Roldana gilgii (Greenm.) H.Rob. & Brettell	1,545,552	28%	1051	99.1%	407	2,105,459	34%	1266	99.5%	11



Tribe	Species name	Accession #
Cardueae	*Arctium lappa* L.	GCA_023525745.1
Vernonieae	*Centrapalus pauciflorus* (Willd.) H.Rob.	GCA_020465985.1
Cichoreae	*Lactuca sativa* L.	GCF_002870075.4
Anthemideae	*Artemisia annua* L.	GCA_014162995.1
Astereae	*Erigeron canadensis* L.	GCF_010389155.1
Heliantheae	*Helianthus annuus* L.	HA412HO v2.0

Figure 1. Phylogeny of Asteraceae tribes and the family's proposed sister group, Calyceraceae, modified from Mandel et al. 2019. Stars at the tip indicate a specimen from that tribe was used for in silico sequencing analyses utilizing CapSim. Colors of stars relate to the table in the bottom left containing sequence accession numbers given by NCBI, excluding *Helianthus annuus* which came from Badouin et al. (2017; https://sunflowergenome.org/assembly-data/).

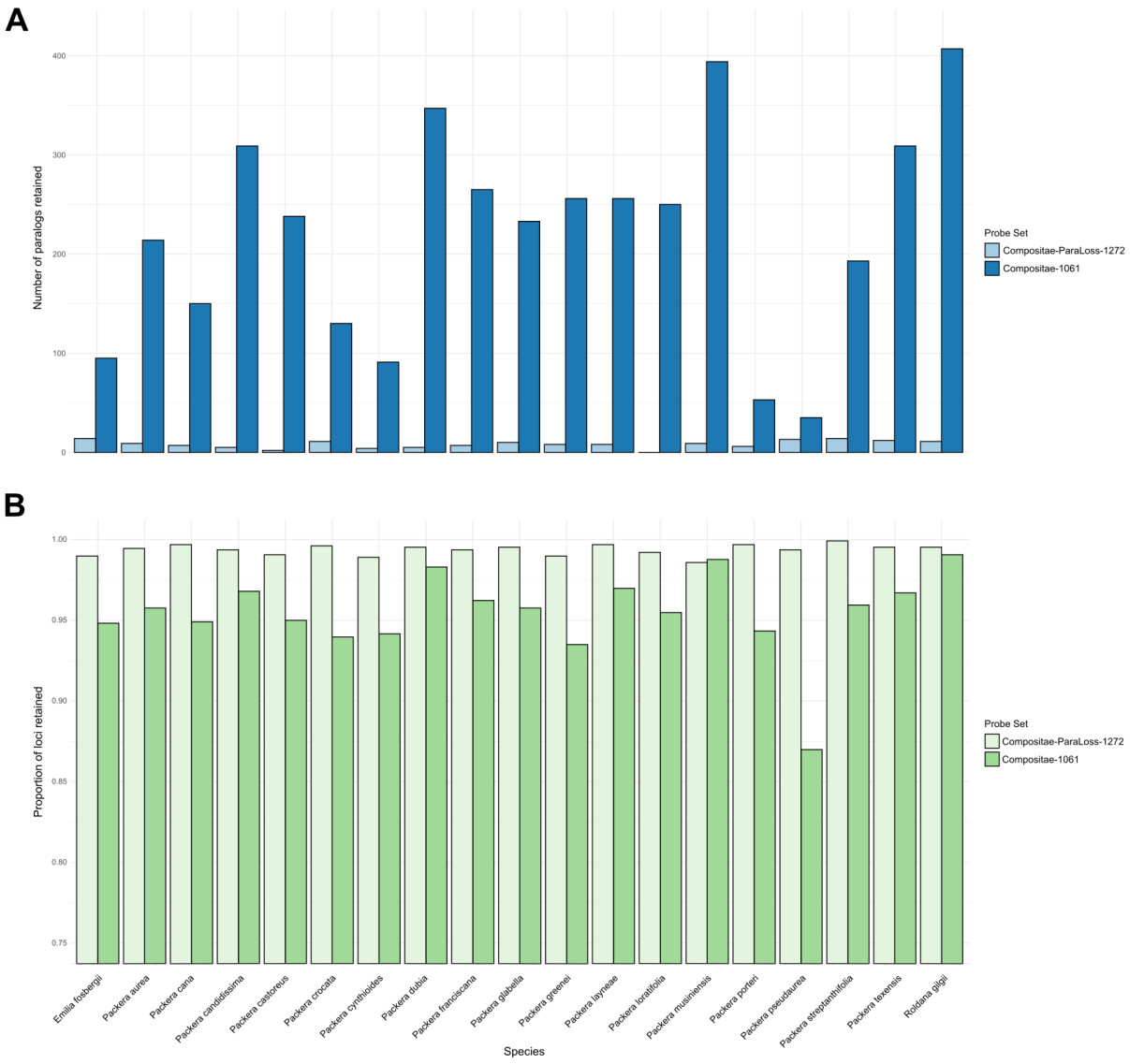


Figure 2. Barplots showing the **A**) number of flagged paralogs, and **B**) the proportion of loci retained for each species dependent on the probe set used. Lighter colors represent the Compositae-ParaLoss-1272 probe set, while darker colors represent the Compositae-1061 probe set as indicated by the keys to the right of the plots. Barplots were generated using base R.

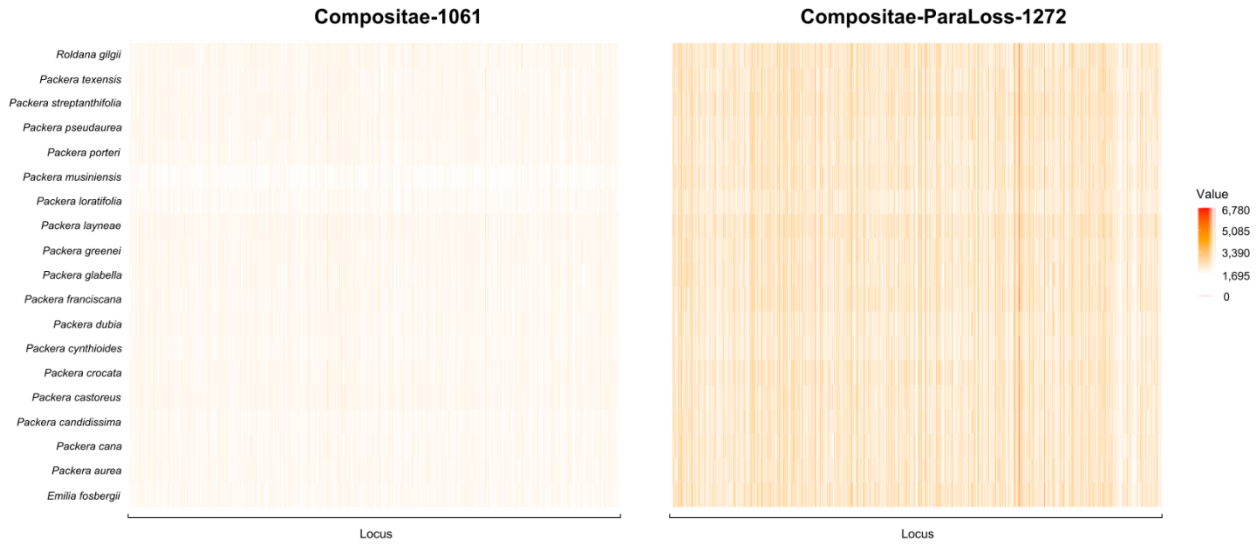


Figure 3. Heatmap of retained locus length in the Compositae-1061 (left) and Compositae-ParaLoss-1272 (right) analyses for each locus (x-axis) of every species (y-axis). The longest loci are indicated by vertical red lines with the smallest loci indicated by vertical orange lines. Loci not retained are shown as white. Heatmaps were generated in R.

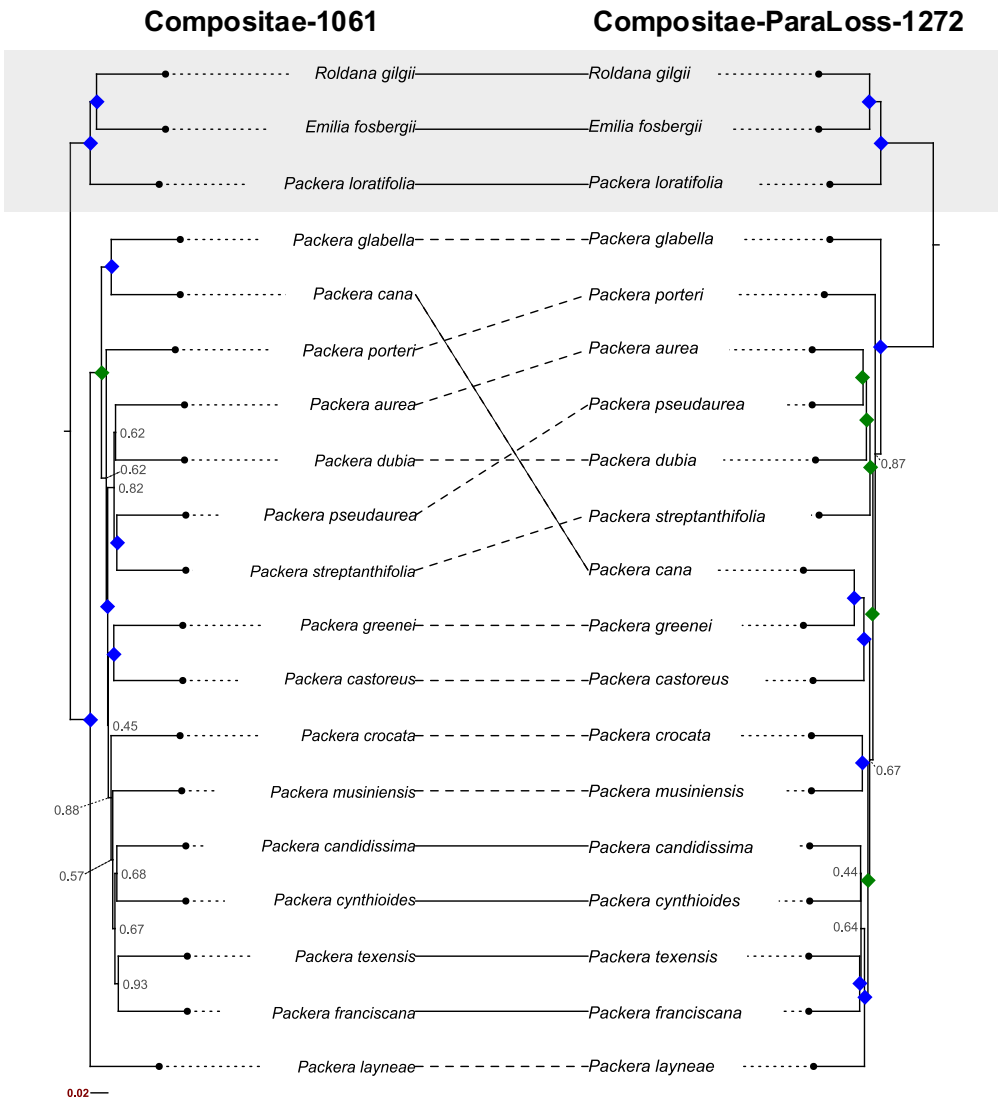


Figure 4. Tanglegram comparing species topologies when phylogenies were developed using the Compositae-1061 probe set (left) or the Compositae-ParaLoss-1272 probe set (right). Topologies representing the same relationship are indicated with a solid line, differing relationships are indicated by a dashed line. Local posterior probability (LPP) values of 1.0LPP are indicated by a blue diamond at the node. LPP values ranging from 0.97-0.99 are indicated by a green diamond. All other LPP values lower than 0.97 are shown at the corresponding node in gray font. Outgroup species are highlighted with a gray shadow box.

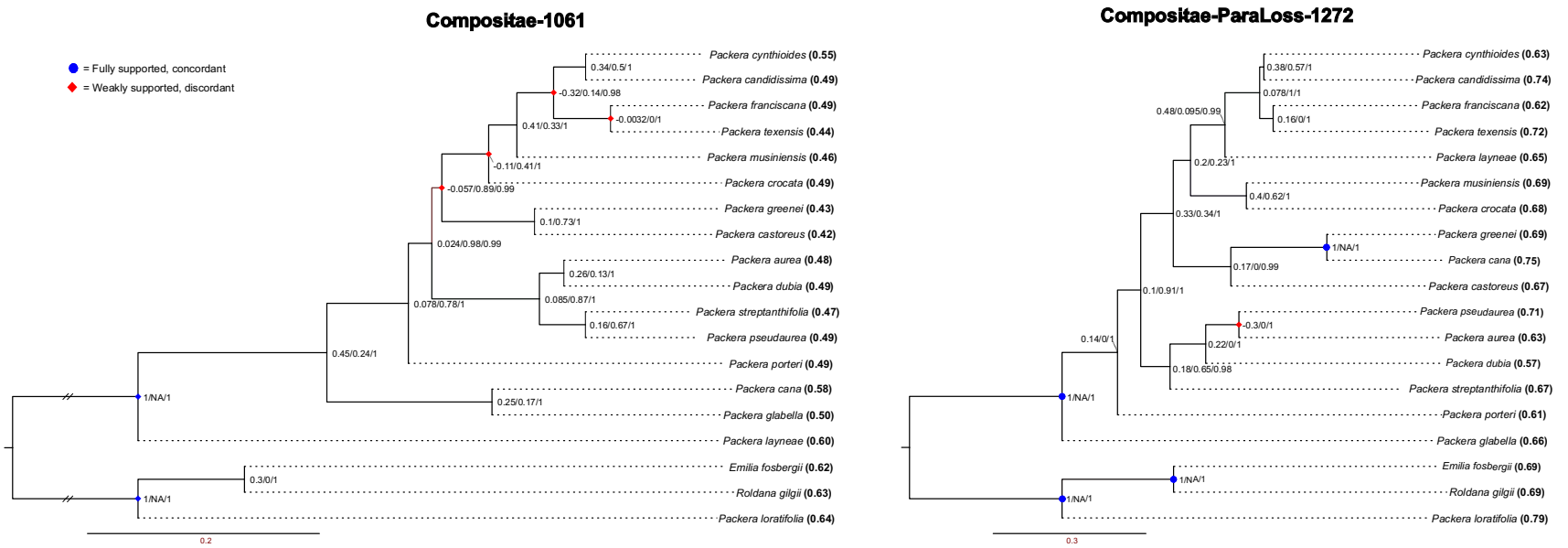


Figure 5. Discordance and support values in the Compositae-1061 (left) and Compositae-ParaLoss-1272 (right) trees indicated by Quartet Sampling. At each node, three values are represented: Quartet Concordance (QC), Quartet Differential (QD), and Quartet Informativeness (QI), shown as QC/QD/QI. Blue circles at the node indicate fully supported and concordant quartets, red diamonds indicate weakly supported and discordant quartets as indicated by Quartet Sampling. Quartet Fidelity (QF) scores are at each tip label in parenthesis and bolded.

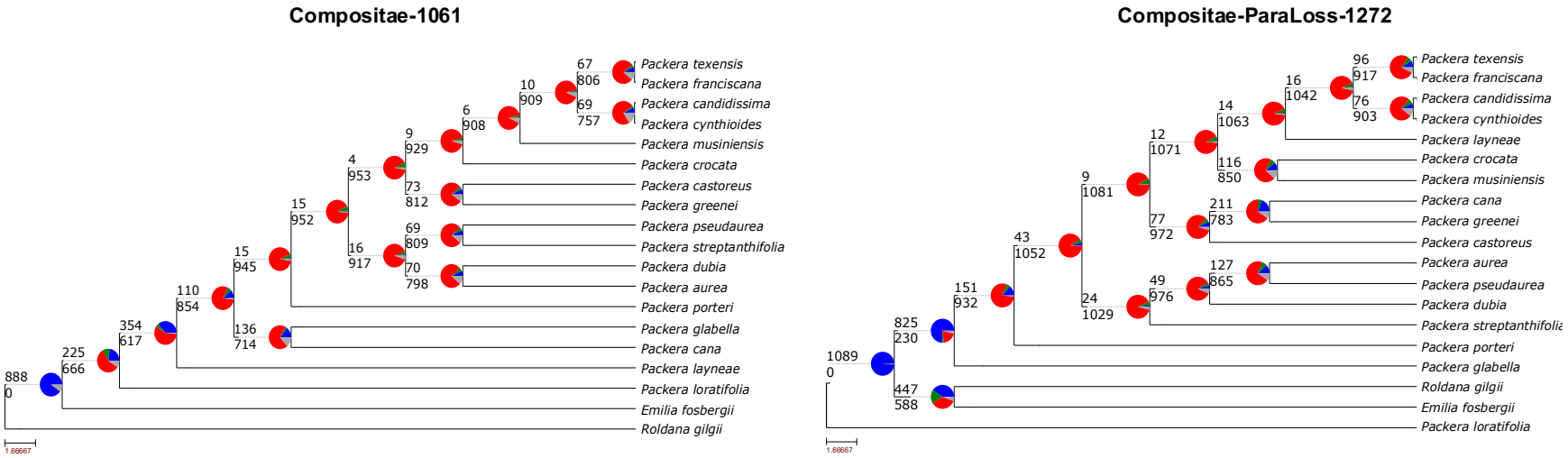


Figure 6. PhyParts results between the Compositae-1061 probe set (left) and Compositae-ParaLoss-1272 probe set (right). Pie charts at nodes show the percentage of gene tree discordance or concordance when compared to the final species tree. The color scheme reveals the percentage of gene trees that are: concordant (blue), the top alternative bipartition (green), all other alternative bipartitions (red), or uninformative at that node (gray). Numbers above and below the branch indicate the number of concordant (blue) and conflicting (red) gene trees, respectively.