Applications in Plant Sciences

APPLICATION ARTICLE

INVITED SPECIAL ARTICLE
For the Special Issue: Exploring Angiosperms353: A Universal Toolkit for Flowering Plant Phylogenomics

Lineage-specific vs. universal: A comparison of the Compositae1061 and Angiosperms353 enrichment panels in the sunflower family

Carolina M. Siniscalchi[1,2,8], Oriane Hidalgo[3,4], Luis Palazzesi[5], Jaume Pellicer[3,4], Lisa Pokorny[3,7], Olivier Maurin[3], Ilia J. Leitch[3], Felix Forest[3], William J. Baker[3], and Jennifer R. Mandel[2,6]

Manuscript received 16 October 2020; revision accepted 15 March 2021.

[1] Department of Biological Sciences, Mississippi State University, Mississippi State, Mississippi 39762, USA

[2] Department of Biological Sciences, University of Memphis, Memphis, Tennessee 38152, USA

[3] Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3AE, United Kingdom

[4] Institut Botànic de Barcelona (IBB, CSIC-Ajuntament de Barcelona), Passeig del Migdia s.n., Barcelona, Catalonia 08038, Spain

[5] División Paleobotánica, Museo Argentino de Ciencias Naturales, CONICET, Buenos Aires C1405DJR, Argentina

[6] Center for Biodiversity, University of Memphis, Memphis, Tennessee 38152, USA

[7] Current address: Centre for Plant Biotechnology and Genomics (CBGP) UPM-INIA, Pozuelo de Alarcón (Madrid) 28223, Spain

[8] Author for correspondence: carol.siniscalchi@gmail.com

Citation: 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(7): e11422.

doi:10.1002/aps3.11422

PREMISE: Phylogenetic studies in the Compositae are challenging due to the sheer size of the family and the challenges they pose for molecular tools, ranging from the genomic impact of polyploid events to their very conserved plastid genomes. The search for better molecular tools for phylogenetic studies led to the development of the family-specific Compositae1061 probe set, as well as the universal Angiosperms353 probe set designed for all flowering plants. In this study, we evaluate the extent to which data generated using the family-specific kit and those obtained with the universal kit can be merged for downstream analyses.

METHODS: We used comparative methods to verify the presence of shared loci between probe sets. Using two sets of eight samples sequenced with Compositae1061 and Angiosperms353, we ran phylogenetic analyses with and without loci flagged as paralogs, a gene tree discordance analysis, and a complementary phylogenetic analysis mixing samples from both sample sets.

RESULTS: Our results show that the Compositae1061 kit provides an average of 721 loci, with 9–46% of them presenting paralogs, while the Angiosperms353 set yields an average of 287 loci, which are less affected by paralogy. Analyses mixing samples from both sets showed that the presence of 30 shared loci in the probe sets allows the combination of data generated in different ways.

DISCUSSION: Combining data generated using different probe sets opens up the possibility of collaborative efforts and shared data within the synantherological community.

KEY WORDS angiosperms; Asteraceae; paralogy; phylogenomics; target capture.

In the past decade, the use of high-throughput sequencing in plant systematics, more specifically Illumina-based short-read sequencing, has changed from being a potentially revolutionary technique to a relatively commonplace approach (Delseni et al., 2010; McKain et al., 2018). While earlier studies relied on genome skimming to obtain organellar (i.e., plastid, mitochondrial, ribosomal) and high-copy-number nuclear markers (Godden et al., 2012; Straub et al., 2012), it soon became clear that for plant systematics, sequencing methods that isolate specific sets of low-copy-number nuclear genomic regions would be more powerful. The first reports of the use of array-free probes for multiplexed in-solution capture and sequencing using high-throughput sequencing platforms were

published in 2011 (Bajgain et al., 2011), followed a few years later by the first family-wide probe set (Mandel et al., 2014).

The Compositae1061 probe set (also known as Compositae COS; Mandel et al., 2014) was one of the first probe sets to be designed specifically for a single family. The sunflower family (Compositae or Asteraceae) comprises more than 25,000 species, and many of its lineages have experienced recent and rapid radiations, large-scale gene family expansions, and ancient polyploidization events (Barker et al., 2008, 2016; Semple and Watanabe, 2009; Huang et al., 2016). Prior to the design and use of the Compositae1061 kit, many of the most important evolutionary questions about the family's diversity were difficult to address, due to the poor resolution of, and lack of support for, the major backbone nodes of the family's phylogeny. With the affordability and efficiency of high-throughput sequencing making genomic approaches attainable in many systems, and under the direction of Compositae expert Vicki A. Funk, members of the synantherology community sought to develop phylogenomic tools to address long-standing evolutionary questions in the family. In early 2011, the Compositae1061 probe set was developed using a set of expressed sequence tag (EST) loci obtained from three economically important members of the sunflower family, lettuce (*Lactuca sativa* L.), safflower (*Carthamus tinctorius* L.), and sunflower (*Helianthus annuus* L.), and included roughly 10,000 probes targeting the exons of 1061 orthologous genes (Mandel et al., 2014).

Since 2014, this probe set has been used to study the family-wide phylogeny (Mandel et al., 2015, 2017, 2019); the relationships among different tribes (Watson et al., 2020); the relationships at the tribe level in the Cardueae (Herrando-Moraira et al., 2018, 2019), Vernonieae (Siniscalchi et al., 2019a), and Perityleae (Lichter-Marck et al., 2020); and the infrageneric relationships in *Antennaria* Gaertn. (Thapa et al., 2020). Data generated using Compositae1061 have also been used as a source to mine for microsatellite markers (Siniscalchi et al., 2019b; Thapa et al., 2019). The utility of the probe set at different evolutionary levels, within the family, and using different starting materials (e.g., herbarium samples vs. samples stored in silica gel) has been extensively explored by Jones et al. (2019). Moreover, the probe set is also able to successfully capture and recover loci for species in families closely related to the Compositae, such as the Calyceraceae and Goodeniaceae (Mandel et al., 2019). Overall, the probe set has been accepted by researchers working on the family, with several ongoing studies yet to be published. In this sense, it fulfills one of the original goals of its design: the creation of a set of markers that could generate easily shareable data across the Compositae.

The wide and varied use of the Compositae1061 probe set has highlighted some of its limitations. One major issue is paralogy (multiple copies of a specific gene), mostly due to the rampant occurrence of both ancient and recent polyploid events within the family (Jones et al., 2019). A second issue is the low phylogenetic resolution at shallow taxonomic levels, such as in studies of closely related taxa or clades resulting from rapid radiation events (Thapa et al., 2020). This issue arises from the probes being designed exclusively from exonic regions, where there might not be enough sequence variation to accurately distinguish the species. Finally, even though the probe set contains 1061 loci, the mean number of loci recovered across studies has been ~700 (Herrando-Moraira et al., 2018).

The recent development of the universal Angiosperms353 kit opens up new opportunities for systematic studies combining deep and shallow phylogenetic levels (Dodsworth et al., 2019). This probe set was specifically developed to choose the minimum number of target instances needed to successfully recover 353 nuclear orthologs from any flowering plant. Its design included 31 Compositae species and a representative of each of the closely related families Goodeniaceae and Menyanthaceae (Johnson et al., 2019). Johnson et al. (2019) showed an average recovery of ~283 loci per species for the Angiosperms353 probe set, and at least 100 loci for over 600 angiosperm species, but this can be increased further using the pipeline recently described by McLay et al. (2021). It has been successfully implemented with low-quality material such as herbarium specimens (Brewer et al., 2019; Shee et al., 2020) and performed well for resolving shallow-level relationships (e.g., radiations [Larridon et al., 2020; Shee et al., 2020] and even at the intraspecific level [Van Andel et al., 2020; Beck et al., 2021; Slimp et al., 2021]). A few studies have compared the performance of Angiosperms353 with other taxon-specific probe sets (e.g., in *Cyperus* L. [Larridon et al., 2020], in the subtribe Malinae of the Rosaceae [Ufimov et al., 2021], and in the Ochnaceae [Shah et al., 2021]), but only Larridon et al. (2020) directly tested the mergeability of different data sets. Alternatively, new lineage-specific kits are being designed that incorporate part or all of the Angiosperms353 targets (e.g., for the Melastomataceae [Jantzen et al., 2020] and the Gesneriaceae [Ogutcen et al., 2021]).

The Angiosperms353 probe set is currently being used in the Plant and Fungal Trees of Life (PAFTOL; http://paftol.org) program at the Royal Botanic Gardens, Kew (Richmond, Surrey, United Kingdom), to produce data for one representative of all angiosperm genera, including Compositae genera. It is also being applied across the Australian flora by the Genomics for Australian Plants consortium (https://www.genomicsforaustralianplants.com/). In this context, a comparison between the Compositae1061 and Angiosperms353 probe sets in the Compositae is timely. Furthermore, understanding how the Angiosperms353 probe set performs in a plant lineage known to contain extensive paralogy issues and how it compares with a family-specific probe set, and verifying if data generated with different probe sets can be combined, is essential in a time where data sharing and collaborative projects abound. Here, we compared the data generated using both probe sets in eight genera of the Compositae. We address the following questions: (1) do the Compositae1061 and Angiosperms353 enrichment panels share any loci?; (2) how do issues of paralogy compare between the two probe sets?; (3) how can we best integrate data generated using these two approaches?

METHODS

Identification of shared loci

The BLAST Command Line Applications were used to examine whether there are any shared loci between the Compositae1061 and Angiosperms353 probe sets. The sequences of the loci contained in the Compositae1061 were used to create a local BLAST database, using the makeblastdb command. As this probe set was based on three EST libraries from sunflower, lettuce, and safflower, some loci are represented by more than one sequence. The sequences of the loci contained in the Angiosperms353 set were obtained from GitHub (https://github.com/mossmatters/Angiosperms353 [accessed 15 April 2020]) and separated into individual FASTA files. The Angiosperms353 probe set contains up to 18 different probe sequences per locus, as it is intended to be applicable across all

flowering plants (details reported by Johnson et al., 2019). Each locus FASTA file separated from the Angiosperms353 set was then queried against the local Compositae1061 BLAST database using BLASTN.

Taxon selection, plant material, DNA extraction, library preparation, and sequencing

Eight taxa were chosen because they overlapped between Mandel et al. (2019) and those available from the PAFTOL program: *Cota tinctoria* (L.) J. Gay (Anthemideae); *Pallenis maritima* (L.) Greuter (Inuleae); *Calendula arvensis* (Vaill.) L. (Calenduleae); *Cardopatium corymbosum* (L.) Pers. (Cardueae); *Cichorium intybus* L. (Cichorieae); *Deinandra corymbosa* (DC.) B. G. Baldwin from PAFTOL and *D. minthornii* (Jeps.) B. G. Baldwin from Mandel et al. (2019) (Heliantheae); *Helichrysum stoechas* (L.) Moench (Gnaphalieae); and *Roldana gilgii* (Greenm.) H. Rob. & Brettell from Mandel et al. (2019) and *R. petasitis* (Sims) H. Rob. & Brettell from PAFTOL (Senecioneae). The taxa sequenced with the Compositae1061 probe set were previously published by Mandel et al. (2019), and details on sample origin and library preparation can be found in Mandel et al. (2014, 2019) and Jones et al. (2019). Sequence data from this project are also available at the National Center for Biotechnology Information Sequence Read Archive under BioProject PRJNA540287. The data set obtained from the Angiosperms353 probe set was collected as part of the PAFTOL program at the Royal Botanic Gardens, Kew, following the protocol described by Johnson et al. (2019), and is available along with voucher information at https://treeoflife.kew.org/ (accessed 4 April 2021).

Sequence assembly and data analysis

Eight samples were assembled for each probe set, as specified above. All sequences were trimmed using Trimmomatic version 0.39 (Bolger et al., 2014), using SLIDINGWINDOW mode with a five-base window and a quality cutoff of 20; reads shorter than 36 bp were removed. The trimmed and paired files were then assembled using HybPiper version 1.3.1 (Johnson et al., 2016), with the respective probe set target sequences as a reference. The trimmed and paired reads were first mapped against the target loci using BWA version 0.7.17 (Li and Durbin, 2009), and were then assembled de novo into contigs using SPAdes version 3.13.1 (Bankevich et al., 2012). Exonerate version 2.2 (Slater and Birney, 2005) was subsequently used to extract the longest unique contig that mapped to a specific target. The final gene matrices were aligned using MAFFT version 7.407 (Katoh and Standley, 2013) with the "-auto" option.

In a second step, each data set was assembled with the opposite probe set reference file to verify the bycatch of loci contained in the other probe set, which was part of our strategy to determine whether data from different sequencing runs generated with different probe sets could be successfully integrated. Even if the two sets contain some of the same loci, the locus lengths can differ depending on the initial source used for probe development. Basic statistics for all assemblies were obtained using the hybpiper_stats.py script in HybPiper and the software AMAS (Borowiec, 2016). Lists of all loci and those loci flagged as paralogous were also obtained with HybPiper. In cases where data generated with one probe set were assembled using the opposite reference file, the recovered loci were further analyzed to identify whether they were in the pool of loci shared by both probe sets, as obtained in the BLAST step described above. All data obtained from these analyses are summarized in Appendices S1–S7 (see Supporting Information).

The recovered loci were used in different phylogenetic analyses. Gene trees were obtained using RAxML version 8.2.9 (Stamatakis, 2014) in the rapid bootstrap mode, with 100 searches. The GTR+I+Γ model was used for all loci, as it is the most complex model currently available and has been shown to accurately infer topologies in real-life and simulated conditions (Abadi et al., 2019). The multispecies pseudo-coalescent method implemented in ASTRAL-III version 5.6.3 (Zhang et al., 2018) was used to obtain a species tree from each data set. Support values in the form of local posterior probabilities were obtained using the "-q" option, and were considered high if equal to or higher than 0.95 and moderate if between 0.90 and 0.94. Four trees were produced: data generated with Compositae1061 and assembled with Compositae1061 (treatment A), data generated with Angiosperms353 and assembled with Angiosperms353 (treatment B), data generated with Compositae1061 and assembled with Angiosperms353 (treatment C), and data generated with Angiosperms353 and assembled with Compositae1061 (treatment D) (Table 1). In a second step, loci flagged as paralogs during the assembly were removed from all four data sets, as defined above, and new gene trees and species trees were obtained with these "cleaned" data sets.

Two additional unrooted ASTRAL trees containing all 16 samples were also produced: one with all 16 samples assembled with the Angiosperms353 probe set as the reference and another with the samples assembled with the Compositae1061 probe set as the reference. A third tree was produced, containing six taxa sequenced with Compositae1061 and two taxa sequenced with Angiosperms353, all assembled using Compositae1061 as the reference and with the loci flagged as paralogs removed, to confirm whether data set integration is indeed possible. All trees were visualized using FigTree version 1.4.4 (https://github.com/rambaut/figtree).

Gene tree discordance was evaluated using PhyParts (Smith et al., 2015). This program requires rooted trees for its analysis; therefore, all four species trees obtained in the ASTRAL analysis, as well as the respective gene trees used as input to generate them, were rooted using the function pxrr in the package phyx (Brown et al., 2017). Due to the lack of an outgroup taxon belonging to a different family, all species trees were rooted with *Cardopatium*, as the tribe Cardueae emerges as sister to the subfamilies Cichorioideae

TABLE 1. Summary of the four treatments used in the study.

Treatment	Sequenced with	Assembled using the reference	No. of samples included	Recovered loci, average (range)	Percent of paralogous loci
A	Compositae1061	Compositae1061	8	721 (3–1012)	0–46%
B	Angiosperms353	Angiosperms353	8	287 (242–323)	0.6–13%
C	Compositae1061	Angiosperms353	6	25 (29–38)	ca. 5%
D	Angiosperms353	Compositae1061	8	35 (21–59)	2–25%

and Asteroideae in most phylogenetic analyses (e.g., Mandel et al., 2019). The gene trees were rooted using a hierarchical scheme (as some missing trees might contain missing taxa) in the following order, after the topology from Mandel et al. (2019): *Cardopatium*, *Cichorium*, *Pallenis*, *Deinandra*, *Calendula*, *Roldana*, *Cota*, and *Helichrysum*. The phypartspiecharts.py script (https://github.com/mossmatters/phyloscripts/tree/master/phypartspiecharts [accessed 15 April 2020]) was used to plot the results from PhyParts as pie charts in each tree node.

RESULTS

Identification of shared loci

The results from the BLAST search show that 59 target instances from Angiosperms353 had hits when queried against the Compositae1061 database. These 59 target instances represent 30 individual loci from the Angiosperms353 probe set, as it contains multiple sequences for each locus. These loci from Angiosperms353 each matched with only one Compositae1061 locus, although some of them had positive hits for more than one probe in the data set, as presented in Table 2. The identity percentage between query and subject sequences varied from 72% to 98% in the searches. Most of the searches generated partial overlaps between the query and the subject, given that locus length is different between each panel. The difference in size for the same loci in each panel varied from 3 to 1836 bp. The results from the searches are summarized in Appendix S1.

TABLE 2. Loci shared across both probe sets. For cases where more than one species representative is included for the same gene, all loci that had hits were included.

Angiosperms353 loci	Compositae1061 loci
Ambtr-6412, NVSO-6412	sunf-At3g23400, saff-At3g23400
Ambtr-6447, IHPC-6447	saff-At2g27290, sunf-At2g27290
Ambtr-6462, EDXZ-6462, LYPZ-6462, NUZN-6462	sunf-At2g27450, saff-At2g27450, lett-At2g27450
AQZD-5614, LVUS-5614, XRCX-5614	sunf-At3g03790, lett-At3g03790
AQZD-5870	sunf-At2g15240
Arath-5477	lett-At1g14620
Arath-5840, EZZT-5840, KDCH-5840, NVSO-5840	lett-At4g35250, saff-At4g35250, sunf-At4g35250
Arath-5857, BXBF-5857, UYED-5857	lett-At2g43030, sunf-At2g43030, saff-At2g43030
AZBL-5841, HYZL-5841, QIKZ-5841, SVVG-5841, UMUL-5841	saff-At1g20540, lett-At1g20540, sunf-At1g20540
BEFC-6449	sunf-At5g57860
BIDT-5562	sunf-At2g36930, lett-At2g36930
BIDT-5910	sunf-At4g25080, lett-At4g25080
BIDT-6733	sunf-At4g37020
BIDT-6946, UMUL-6946, VUSY-6946	sunf-At3g03100, saff-At3g03100, lett-At3g03100
BIDT-6954, IDAU-6954	sunf-At1g50575
DOVJ-7371	lett-At4g13500
DUNJ-6498, HLJG-6498, OXYP-6498, RCUX-6498	sunf-At1g55670, lett-At1g55670
EMBR-5918, JYMN-5918	saff-At2g31040, sunf-At2g31040, lett-At2g31040
HOKG-6458	lett-At1g04620
JEPE-4527	sunf-At1g09830
JNKW-6705	lett-At3g62810
JYMN-7141	sunf-At3g55250
LSJW-5933, MDJK-5933	sunf-At3g63140, lett-At3g63140
NUZN-6139	lett-At1g75330
NVSO-7194	sunf-At1g76450
Orysa-6038	saff-At4g32770
QIKZ-7367, WAIL-7367, ZCUA-7367	sunf-At2g03420
QTJY-6068, UYED-6068	sunf-At3g05070, saff-At3g05070, lett-At3g05070
VVPY-6913, WYIG-6913	lett-At5g23120
XRCX-5594	lett-At1g76080

Recovered loci and paralogy

The main results from the assembly are summarized in Fig. 1, Table 3, and Appendices S2 and S3. Alignments produced from data sequenced with Angiosperms353 and assembled with the matching reference (treatment B) tended to be longer and have more parsimony-informative (PI) sites (Fig. 1A), while those sequenced and assembled with Compositae1061 (treatment A) tended to have similar lengths. The assemblies with the opposite reference (treatments C and D) tended to produce shorter alignments, and in the case of treatment D, several alignments did not present PI sites. For the samples sequenced with the Compositae1061 probe set and assembled using the same probe set as reference (treatment A), the percentage of reads on target varied from 2.8% in *Cichorium* to 56% in *Helichrysum* (Fig. 1B). The number of recovered loci varied from three in *Cichorium* to 1012 in *Deinandra*, with an average of 721 loci recovered. Loci flagged as paralogous were recovered for six of the eight species and the percentage of paralogous loci in relation to the recovered loci varied from 9% (*Cardopatium*) to 47% (*Calendula*). When the same data set was assembled with the Angiosperms353 probe set (treatment C), it showed percentage of reads on target ranged from 0.1% (*Cichorium*) to 1.6% (*Pallenis*). The number of recovered loci varied from 28 (*Cardopatium*) to 38 (*Calendula*). Only *Calendula* and *Helichrysum* presented paralogs, both with around 5% of the recovered loci being flagged. *Cichorium* and *Roldana* had very few loci recovered, being dropped from the final assembly with Angiosperms353, which was probably due to issues during genomic library preparation, well before sequencing.

For the samples sequenced with the Angiosperms353 probe set and assembled using this probe set as reference (treatment B), the percentage of reads on target varied from 9.5% (*Cota*) to 18.4% (*Calendula*) (Fig. 1B) and the number of recovered loci was somewhere between 242 (*Cota*) and 323 (*Pallenis*), with an average of 287 loci. The percentage of loci flagged as paralogous ranged from 0.3% (*Pallenis*) to 13% (*Calendula*). When this data set was assembled with the Compositae1061 probe reference (treatment D), the percentage of reads on target varied from 1.8% (*Cota*) to 5.2% (*Calendula*). The number of recovered loci varied from 21 (*Cota*) to 59 (*Calendula*), with an average of 32 loci. The percentage of paralogous loci varied from 2% (*Pallenis*) to 25% (*Calendula*).

The assembly of data generated in treatment C generated 39 unique loci. From these 39 loci, 29 are contained in the pool of 30 loci that are represented in both probe sets. The opposite scenario, treatment D, generated 71 unique loci, among which all 30 loci shared by both probe sets are represented (Appendix S4).

In treatment A, 640 of the 1061 loci (~60%) that compose the probe set were flagged as paralogous during assembly with HybPiper. Of these 640 loci, 388 were flagged for two or more taxa, while the remaining 252 loci were flagged only in one taxon (Appendix S5).

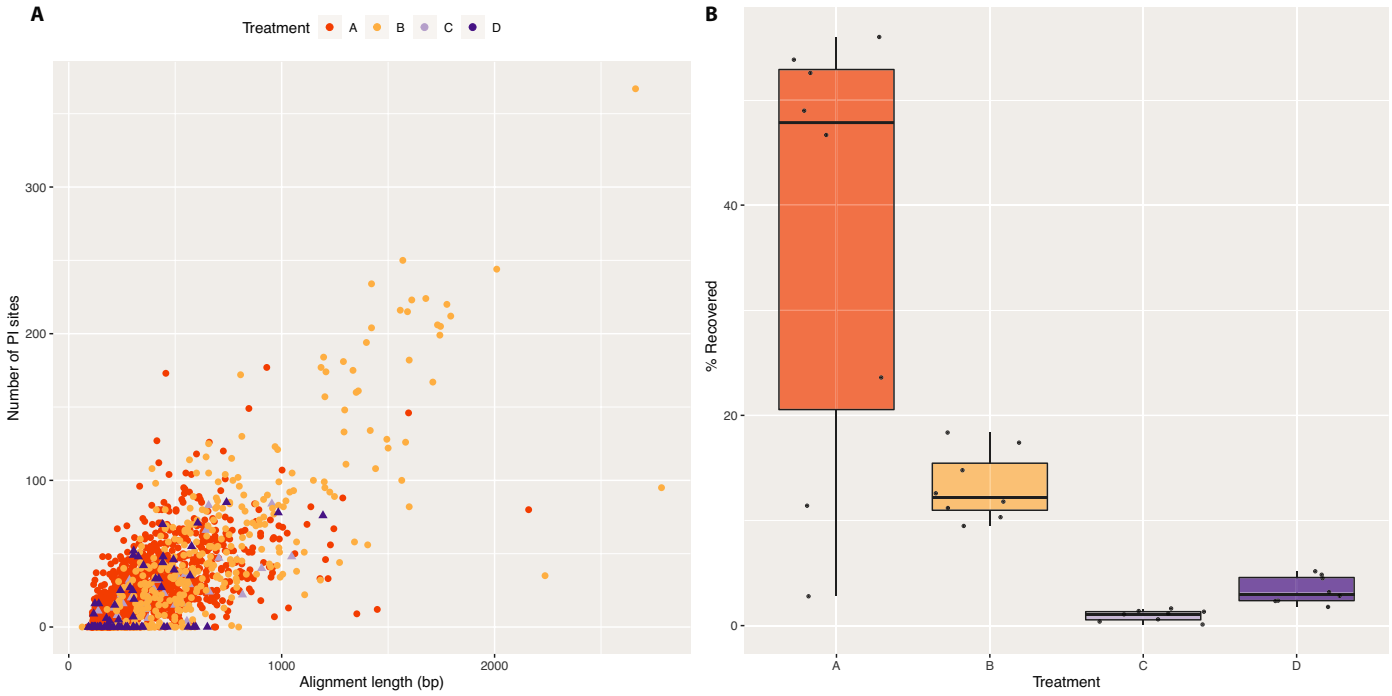

FIGURE 1. Basic assembly statistics. (A) Number of parsimony-informative (PI) sites in relation to the alignment length. Circles represent data sequenced and assembled using the same probe set as a reference, while triangles represent an assembly using the other probe set as a reference. (B) Percentage of reads mapping to targets (recovered) in each treatment. Error bars represent the 25th and 75th percentiles.

For treatment B, 16 loci were flagged as paralogous in two or more samples and 43 in only one sample (Appendix S6), totaling 58 loci (16% of the total in the probe set). Paralogous loci recovered from the sequences assembled with the opposite reference (treatments C and D) are summarized in Appendix S7.

Phylogenetic relationships and gene tree discordance

The recovered phylogenetic relationships varied depending on the data set used to generate them (Fig. 2). The four trees were rooted using *Cardopatium* (Cardueae). All eight samples were recovered in three of the trees, except *Roldana* and *Cichorium* in the tree based on treatment C. In all three completely sampled trees, *Cichorium* (Cichorieae) was sister to the larger subfamily Asteroidae clade. Within this clade, *Deinandra* (Heliantheae) and *Pallenis* (Inuleae) were always sister taxa.

The topologies of the trees obtained from the Compositae1061 data were similar regardless of whether the data were assembled with Compositae1061 (Fig. 2A) or Angiosperms353 (Fig. 2C) as a reference. In root-to-tip order, *Calendula* was in a grade leading to a *Cota*–*Helichrysum* clade. *Roldana* was sister to this grade in treatment A. In the trees generated in treatment D (Fig. 2D), *Calendula*–*Roldana* and *Cota*–*Helichrysum* formed two sister clades. In the tree resulting from treatment B (Fig. 2B), *Calendula* was sister to a five-species clade, in which *Deinandra* and *Pallenis* formed a clade sister to *Helichrysum* and the *Roldana*–*Cota* clade. In the tree based on treatment C, in which *Cichorium* and *Roldana* were not recovered (Fig. 2C), a *Deinandra*–*Pallenis* clade was sister to a clade formed by *Calendula* and a *Helichrysum*–*Cota* clade.

The removal of paralogous loci resulted in topological changes in half of the trees (Fig. 3). In the tree obtained from treatment A (Fig. 3A), *Cichorium* was dropped and *Roldana* emerged in a clade with *Deinandra* and *Pallenis* instead of grouping with the other three species. *Calendula* emerged as sister to *Cota*, a relationship not seen in other trees. In the tree obtained from treatment D (Fig. 3D), *Cichorium* emerged within the Asteroideae clade, although with very low support. In the two trees that did not present topological changes (Fig. 3B, C), slight changes in support were observed, with the treatment B tree presenting decreased support in the backbone but not in the internal nodes and treatment C presenting the opposite pattern.

In the two trees obtained with the complete 16-species set assembled either with Compositae1061 or Angiosperms353 as the reference (Fig. 4), all species pairs form individual clades, except for *Cichorium* and *Roldana* in the Compositae1061 tree (Fig. 4B); these two species were not recovered from the data generated with Angiosperms353. In the tree obtained from the mix of both data sets (Fig. 5), the two samples sequenced with Angiosperms353 emerged in expected positions, with *Cichorium* as sister to the Asteroideae clade, and *Roldana* in the clade with *Cota*, *Helichrysum*, and *Calendula*, although the relationships in this four-species clade are different from those seen in other topologies.

The results of the gene tree discordance analyses (Fig. 6) show a panorama of wide disagreement, increasing toward the tips of the trees. In the tree for treatment A (Fig. 6A), the percentage of gene trees agreeing with the species tree at each node varied from 0.5% to 100%, with most of the nodes presenting values around 20%. For treatment B (Fig. 6B), this percentage varied from 10% to 100%, with most nodes in the 10% to 20% range. Gene tree discordance was lower in the trees generated from data assembled with the opposite reference (Fig. 6C, D), with the percentage of concordant gene trees staying around 30% in most nodes in both trees.

TABLE 3. Summary of assembly statistics.

Treatment	Sample	Genes recovered	Genes flagged as paralogs	Genes not flagged as paralogs	Percentage of genes flagged as paralogs
A (data generated with Compositae1061 and assembled using Compositae1061 as the reference)	*Calendula*	1008	470	538	47%
	Cardopatium	893	82	811	9%
	Cichorium	3	0	3	0
	Cota	903	211	692	23%
	Deinandra	1012	301	711	29%
	Helichrysum	951	248	703	26%
	Pallenis	977	181	796	18%
	Roldana	23	0	23	0
B (data generated with Angiosperms353 and assembled using Angiosperms353 as the reference)	*Calendula*	315	41	274	13%
	Cardopatium	314	2	312	0.6%
	Cichorium	296	3	293	1%
	Cota	242	5	237	2%
	Deinandra	272	5	267	2%
	Helichrysum	275	6	269	2%
	Pallenis	323	1	322	0.3%
	Roldana	261	11	250	4%
C (data generated with Compositae1061 and assembled using Angiosperms353 as the reference)	*Calendula*	38	2	36	5%
	Cardopatium	29	0	29	0
	Cichorium	0	0	0	NA
	Cota	31	0	31	0
	Deinandra	37	0	37	0
	Helichrysum	35	2	33	5%
	Pallenis	35	0	35	0
	Roldana	0	0	0	NA
D (data generated with Angiosperms353 and assembled using Compositae1061 as the reference)	*Calendula*	59	15	44	25%
	Cardopatium	34	0	34	0
	Cichorium	31	0	31	0
	Cota	21	0	21	0
	Deinandra	31	0	31	0
	Helichrysum	30	2	28	6%
	Pallenis	48	1	47	2%
	Roldana	28	1	27	3%

DISCUSSION

In the present study, we sought to compare two different enrichment panels: Compositae1061, developed based on genomic resources available only for the sunflower family (EST libraries), and Angiosperms353, designed from angiosperm-wide genomic resources (transcriptomes and genomes). One of the goals of the present study was to verify the presence of shared loci in both sets. We identified 30 loci that are included in both probe sets, which facilitate complementary analyses with data combined from different studies. These loci appear to be among those that are consistently recovered across the family, given that assembling data generated with one of the probe sets with the opposite reference resulted in the recovery of similar numbers of genes across the samples, as seen in Table 3. Additionally, as seen in Fig. 4, the variation present in the limited number of shared loci is sufficient to allow for samples of the same taxa, sequenced with different probe sets, to group together in phylogenetic analysis. Mixing samples obtained with different probe sets, but assembled with the same reference, also proved possible (Fig. 5), with similar topology and support values seen in the trees generated using the data assembled with their own references.

The varying phylogenetic relationships we recovered are representative of the issues routinely found during phylogenetic studies in the sunflower family. From the eight sampled species, six belong to the subfamily Asteroideae, which includes >60% of the species diversity of the family (Susanna et al., 2020). Relationships among groups of tribes in this subfamily have been notoriously difficult to resolve, such as the tribes within the Heliantheae alliance and the group informally named Fab5 (Anthemideae, Astereae, Calenduleae, Gnaphalieae, and Senecioneae), as seen in multiple Compositae phylogenetic studies (Pelser and Watson, 2009; Huang et al., 2016; Panero and Crozier, 2016; Mandel et al., 2019; Watson et al., 2020). In the trees presented in Figs. 2 and 3, the conflict among species belonging to the Fab5 (*Cota*, *Calendula*, *Helichrysum*, and *Roldana*) is clear, with their relationships changing in each tree and with the removal of paralogs. It is noteworthy that relationships recovered with the Compositae1061 data set do not reflect those shown in Mandel et al. (2019), although they are based on the same data, suggesting that the reduced sampling used in this study significantly impacted the resolution of the relationships. The low number of loci recovered for *Cichorium* (3) and *Roldana* (23) in this data set, likely a result of sequencing issues, could be an additional source of phylogenetic noise and a factor leading to topological incongruences.

With the increasing abundance of large-scale genomic data sets composed mainly of nuclear genes, the issue of paralogy (which results from small-scale gene family expansions to whole-genome duplications) has become more widely discussed among plant systematists. Multiple copies of specific genes or whole gene families are most likely a consequence of the ancient polyploid origin at the base of all flowering plants and the occurrence of further additional ancient polyploid events leading to the base of the Compositae

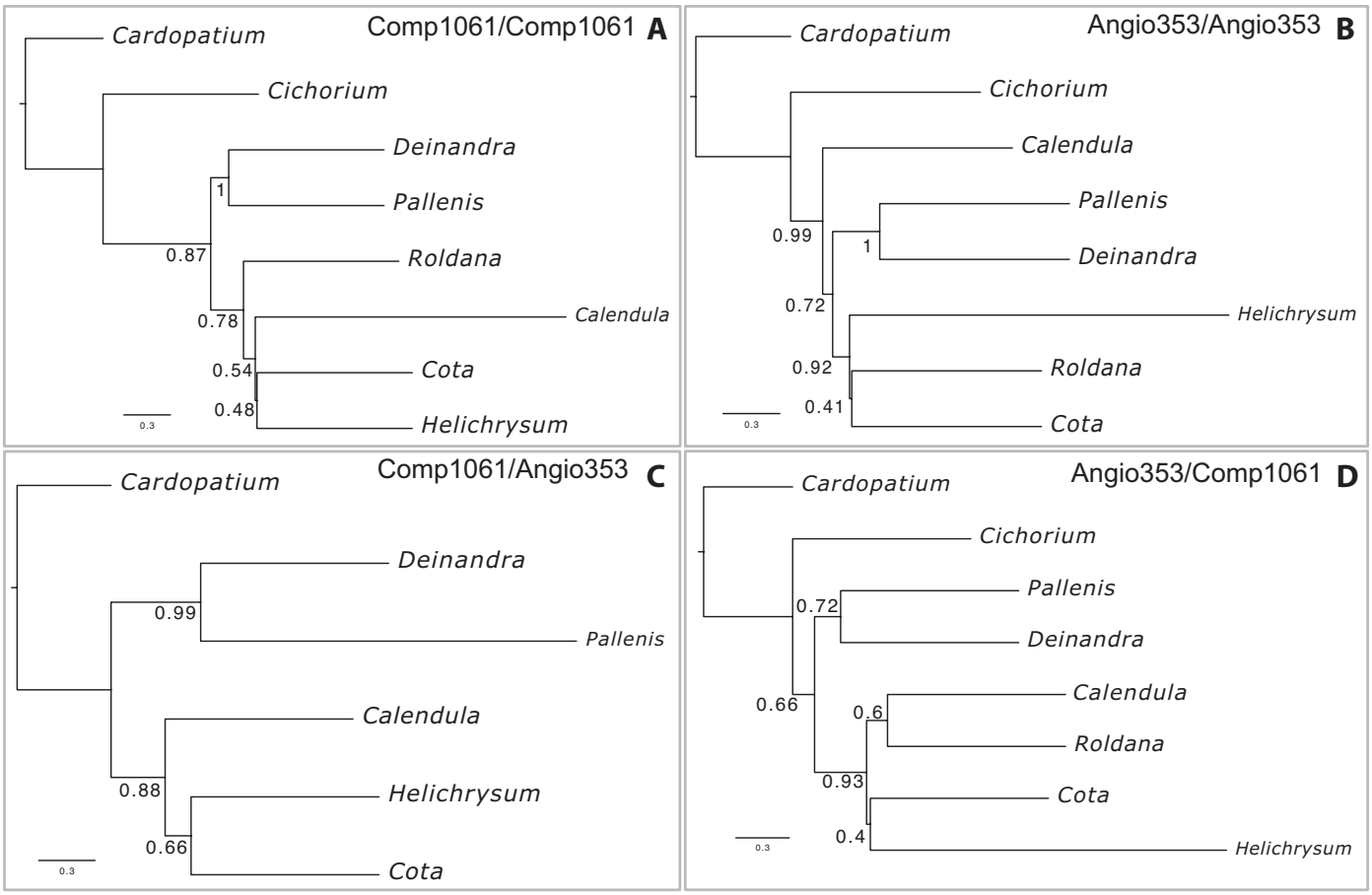

FIGURE 2. Phylogenies obtained using the different data sets and assembly strategies. Values on the nodes are local posterior probabilities obtained using ASTRAL-III. (A) Data generated with Compositae1061 and assembled using Compositae1061 as the reference. (B) Data generated with Angiosperms353 and assembled using Angiosperms353 as the reference. (C) Data generated with Compositae1061 and assembled using Angiosperms353 as the reference. (D) Data generated with Angiosperms353 and assembled using Compositae1061 as the reference.

(Wendel, 2015; Van de Peer et al., 2017; Leebens-Mack et al., 2019). Indeed, a hexaploid ancestor was proposed for most of the lineages within the family (Barker et al., 2016; Li and Barker, 2020), and paralogy has been a frequent issue in phylogenomic studies ever since their inception.

The design of the Compositae1061 probe set focused on genes considered to be conserved orthologs at the time, thus aiming to reduce the number of possibly paralogous genes in the set (Mandel et al., 2014); however, with its use across different lineages of the family, it became clear that large numbers of loci are present in multiple copies after sequencing. As previously demonstrated by Jones et al. (2019), lineages within the family present different degrees of paralogy, probably associated with the hypothesized presence of further whole genome duplication events in several lineages (Huang et al., 2016; Li and Barker, 2020). Corroborating the results from Herrando-Moraira et al. (2018) and Jones et al. (2019), for the data generated using the Compositae1061 probe set, ~20% of the loci are putatively paralogous in most species, while *Cardopatium* (Cardueae) presented the lowest levels of paralogy (~9%). There is currently no evidence for ancient polyploidy in the tribe Cardueae, while there are multiple events proposed for the other tribes present in our analysis, such as the Calenduleae (*Calendula*), the Gnaphalieae (*Helichrysum*), and the Heliantheae alliance (*Deinandra*) (Huang et al., 2016; Li and Barker, 2020). As expected, the removal of paralogs from the analysis caused topological changes (Fig. 2A, 3A), with more marked effects in the position of taxa with higher paralogy, such as *Calendula* and *Helichrysum*.

The number of paralogous loci recovered from the Angiosperms353 data assembled with itself as the reference reflect these phylogenetic trends, with *Calendula* presenting the highest proportion of paralogs (~13%) and *Cardopatium* presenting the second smallest (~0.6%). *Pallenis* (Inuleae) is the sample with the smallest number of paralogs (0.3%) in this set. The very high paralogy observed in *Calendula arvensis* in both data sets likely arises from the fact that the species has not only experienced multiple ancient polyploid events, but is also an allotetraploid (Nora et al., 2013; Plume, 2015). The overall lower levels of paralogy seen in the Angiosperms353 data set appear not to interfere with the topology as there are no changes when paralogs are removed, although this generates a slight improvement in support values (Figs. 2B, 3B).

Regarding the assemblies carried out using the other probe set as the reference, in the data generated for treatment C (Fig. 2C), the number of paralogous loci in each species decreased in relation to the number generated for treatment A, while in treatment D (Fig. 2D), the number of paralogous loci increased for the four species (Appendix S7). The treatment D assembly recovered 20 unique loci flagged as paralogs, of which 11 were from the pool of 30 loci

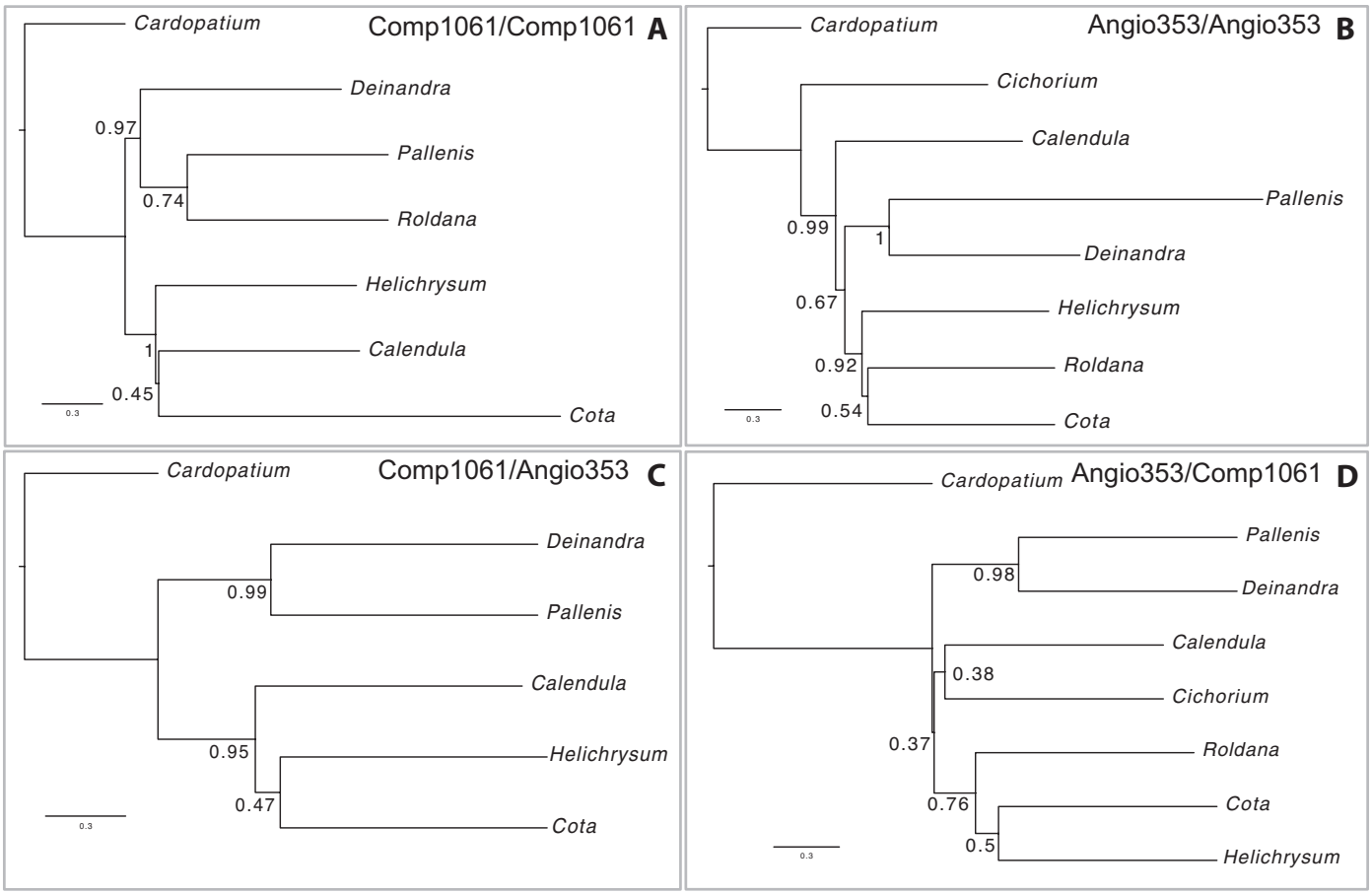

FIGURE 3. Phylogenies obtained using the different data sets and assembly strategies after the removal of loci flagged as paralogs. Values on the nodes are local posterior probabilities obtained using ASTRAL-III. (A) Data generated with Compositae1061 and assembled using Compositae1061 as the reference. (B) Data generated with Angiosperms353 and assembled using Angiosperms353 as the reference. (C) Data generated with Compositae1061 and assembled using Angiosperms353 as the reference. (D) Data generated with Angiosperms353 and assembled using Compositae1061 as the reference.

shared by both probe sets. The recovery of fewer paralogous loci when assembling Compositae1061 data with the Angiosperms353 reference might be explained by an overall lower rate of potentially paralogous loci in the Angiosperms353 kit (see further discussion of this below). The removal of paralogs from these two assemblies caused different effects: the topology remained the same in the first tree, with small changes in support (Fig. 3C), but was altered in the second case (Fig. 3D). The position of *Cichorium* in this last topology, within the subfamily Asteroideae, is dubious and goes against all previous phylogenetic work and the historical classifications of the Compositae. *Calendula* was the taxon with the highest number of removed loci, emerging as sister to *Cichorium*, and is likely the reason for the stark topological changes.

Samples sequenced with Angiosperms353 present a much smaller proportion of paralogs than those sequenced with Compositae1061, being below 10% in seven of the eight species. This is probably a reflection of the original data used to develop each probe set. While the Compositae1061 set was based exclusively on three EST libraries, the only genetic resources available for the Compositae at the time, the development of the Angiosperms353 probe set relied on a set of 410 alignments of orthologous loci across 1100 green plants, singled out in the context of the 1000 Plants (1KP) project (details in Johnson et al., 2019; Leebens-Mack et al., 2019). This comparison across a wide group of genomic references, including 31 Compositae species, allows for a more refined selection of target regions that will truly present as single-copy loci in most plant species. This data set also seems to be more resilient to paralogy, as there are no changes in topology with the removal of paralogous loci. It is worth noting, however, that even if the loci flagged as paralogous were removed from the analysis, the Compositae1061 set still generates several hundred more loci than Angiosperms353 (538–811 vs. 237–322, respectively) and presents higher proportions of on-target reads (Fig. 1B), which could be decisive when dealing with rapid radiations or very recent divergences. Nevertheless, removing the paralogs created changes in topology in the present study, which could also be a consequence of the very sparse sampling.

Many phylogenomic studies using the Compositae1061 probe set assembled their data using phyluce (Faircloth et al., 2015), which takes a more restrictive approach than HybPiper with regard to paralogs, as a way of dealing with resulting conflicting relationships (Mandel et al., 2019; Siniscalchi et al., 2019a; Thapa et al., 2019). HybPiper flags possible paralogs but keeps the locus in the final alignments by choosing either the copy with the greater sequencing depth or the one with the greatest percentage identity to the reference (Johnson et al., 2016). On the other hand, phyluce removes any locus for which the assembled contigs match multiple loci or

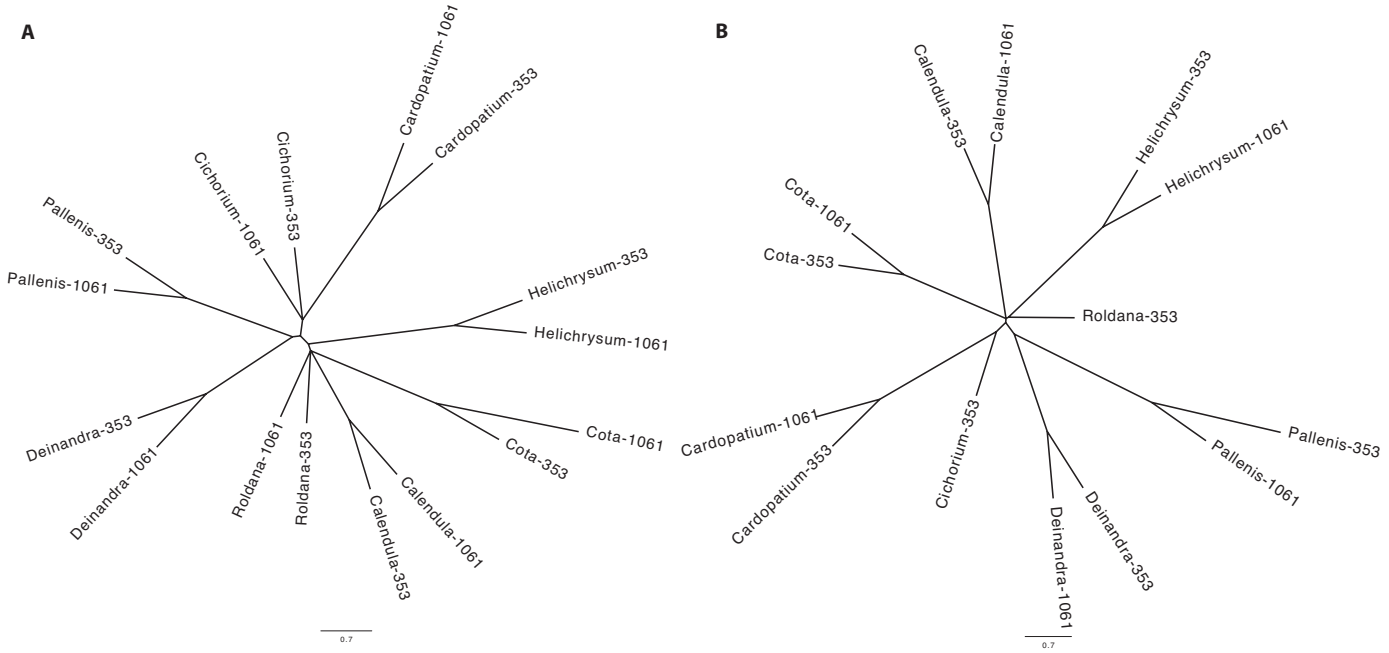

FIGURE 4. Phylogenies combining all 16 samples. (A) All samples assembled using Compositae1061 as the reference. (B) All samples assembled using Angiosperms353 as the reference. The suffixed numbers refer to the probe set used.

different contigs match the same locus from the final alignments. One additional difference between pipelines is also the extent of the assembled loci. The first step in HybPiper is mapping against reference loci (target sequences), followed by a de novo assembly (of mapped reads per target locus); thus, the loci ultimately recovered usually span the length of the reference loci and include flanking regions, which are later automatically removed and can be recovered using the "-supercontig" option in the "retrieve_sequences.py" script. Phyluce begins with a de novo assembly of the data, with the contigs being posteriorly matched against the references, which also allows for the assembly of off-target flanking regions or introns, which are not removed. Herrando-Moraira et al. (2018) compared the effects of different assembly methods in the relationships of the tribe Cardueae and noted that phyluce introduces more phylogenetic noise, but without deeply affecting the recovered relationships. This is likely an effect of the unequal recovery of flanking regions; as they are not targeted, the recovery is different across loci and taxa, introducing more missing data in the final matrices.

In the present study, we decided to assemble the data with HybPiper, as it is more widely used by the plant systematics community. We analyzed data sets with and without paralogs. Given the widespread genomic duplications in the family, it is prudent to remove potentially paralogous loci from the final assemblies, or at least investigate the phylogenetic history of possible paralogous copies, for example, using the "paralog_investigator.py" script in HybPiper. Unfortunately, HybPiper does not include an option to easily remove specific paralogs from samples, in which case phyluce tends to be a better option, as this is done automatically in its pipeline. Finally, as HybPiper assemblies tend to be similar in length to the original locus sequences, the sequences and assemblies generated using Angiosperms353 in this study are longer overall than those from Compositae1061 (Fig. 1A), which is probably due to the original size of the loci used as basis to design the probes contained in each kit.

Another issue arising from large phylogenomic data sets becoming more widely available is gene tree discordance, which is usually explained by whole-genome duplication or polyploidy, hybridization, incomplete lineage sorting, or some combination of these processes. Gene tree discordance has been widely documented in plants (as summarized by Smith et al., 2020) and more specifically in the Compositae (Herrando-Moraira et al., 2019; Jones et al., 2019; Siniscalchi et al., 2019a; Watson et al., 2020). Most gene tree discordance analyses in Compositae studies show high levels of disagreement, increasing toward the tips of the trees. This is likely due to the fact that gene recovery is variable in samples sequenced with the Compositae1061 probe set, an issue that might have several origins, such as low probe hybridization efficiency, large and repetitive genomes impairing probe binding, and a high divergence of the probes in relation to the target sequence. However, one study in the Cardueae (Herrando-Moraira et al., 2019) showed the opposite, with the backbone presenting more discordance than the tips. Siniscalchi et al. (2019a) demonstrated that reduced data sets, which eliminated gene trees lacking several taxa, improved the overall discordance by decreasing the number of uninformative gene trees and increasing the proportion of gene trees that agree with the species tree. A similar effect is observed in the discordance analysis presented here, where the two trees obtained with data assembled with the opposite probe set as the reference showed less overall discordance, probably due to the lower number of missing taxa in these gene trees, as presented in Appendix S2.

Overall, our results show that the Compositae data obtained using two different probe sets can be combined due to the presence of 30 shared loci between them, enabling mixed analyses (Figs. 4, 5). One interesting result is that both data sets assembled using the other as a reference recovered more than 30 loci each. Hybridization reactions vary in precision and efficiency, with the occurrence of bycatch being well known (e.g., Jones et al., 2019). One explanation could be the presence of other loci contained in the probe set being part of the bycatch of the hybridization, which then match to targets

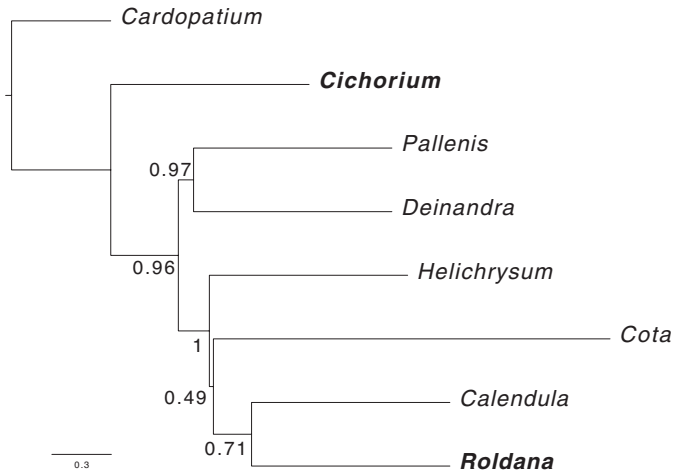

FIGURE 5. Phylogeny combining six samples sequenced with Compositae1061 (regular text) and two with Angiosperms353 (bold text), all assembled using Compositae1061 as the reference. Node values are local posterior probabilities.

on the reference file. However, our BLAST-based approach to match both probe sets could have been too stringent, not capturing loci that were too divergent or that had short overlaps between them.

Few studies have used Compositae1061 to investigate infrageneric relationships in different Compositae tribes, while there are currently no published reports of the use of Angiosperms353 in the Compositae. Lichter-Marck et al. (2020) and Thapa et al. (2020) investigated infrageneric relationships within *Perityle* Benth. (Perityleae) and *Antennaria* (Gnaphalieae), respectively, and compared the effects of concatenated vs. gene tree analyses. Both studies found high levels of paralogy and topological incongruence between the phylogenies generated using different inference methods. Jones et al. (2019) investigated the levels of paralogy and conflict within a species complex in *Picris* L. (Cichorieae), showing high levels of gene tree conflict but good overall resolution and support. The probe set has been proven useful at lower phylogenetic levels but presented the same issues seen at higher levels, with incongruences between different assembly and phylogenetic inference methods (Herrando-Moraira et al., 2018; Siniscalchi et al., 2019a), possibly indicating issues with the actual loci chosen as targets or a complicated history of genomic evolution within the family.

The possibility of integrating data from different origins opens up opportunities for new collaborations and integrative projects using data that can be universally shared. Given the high amount of paralogy in the loci contained in the Compositae1061 set and taking into account the new genetic resources available for the family, such as three complete genomes and more than 30 transcriptomes, a redesign of this specific probe set could be beneficial. A deeper study of

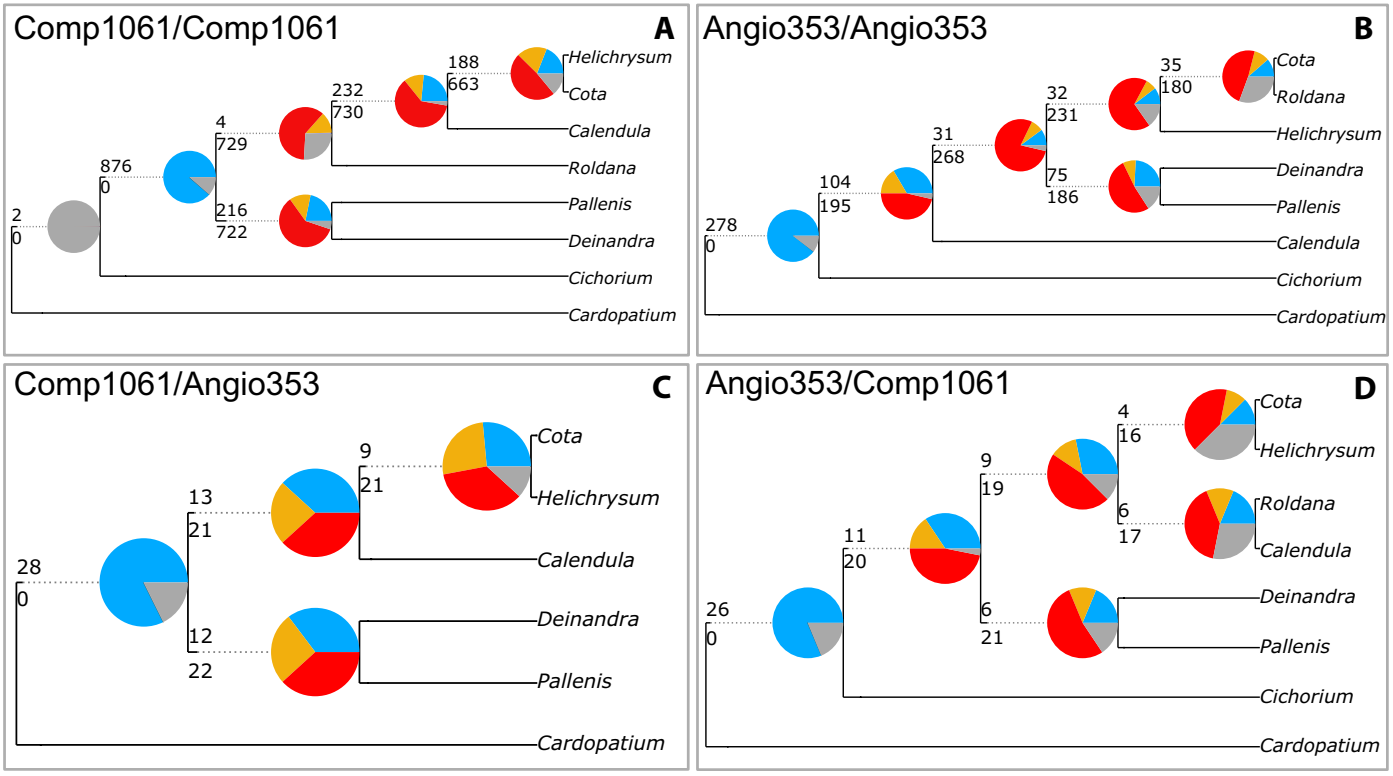

FIGURE 6. Gene tree discordance analysis. Pie charts represent the proportion of gene trees that support that specific node. Blue represents gene trees agreeing with the species tree, orange those that agree with the main alternative topology, red those that agree with all other topologies, and gray the proportion of uninformative trees. The numbers on the branches represent the number of concordant gene trees (top) and the number of conflicting trees (bottom). (A) Data generated with Compositae1061 and assembled using Compositae1061 as the reference. (B) Data generated with Angiosperms353 and assembled using Angiosperms353 as the reference. (C) Data generated with Compositae1061 and assembled using Angiosperms353 as the reference. (D) Data generated with Angiosperms353 and assembled using Compositae1061 as the reference.

paralogy across the family could indicate loci that are problematic in several lineages, and these could then be replaced by newly selected ones. Alternatively, the loci contained in the Angiosperms353 set could be included with the Compositae1061 to create a more inclusive set of targeted loci, which has already been done for the Melastomataceae (Jantzen et al., 2020) and Gesneriaceae (Ogutcen et al., 2021). Finally, it is worth noting that Mandel et al. (2019) successfully integrated transcriptomic data from the 1KP project with the Compositae1061 loci, demonstrating how different sources of data can be combined for phylogenetic reconstruction.

This data integration will be useful at higher levels of phylogenetic analyses, such as for adding outgroups to an analysis or in tribe or subfamily phylogenies, as the small number of shared loci between Compositae1061 and Angiosperms353 will probably not be sufficient to resolve relationships in shallower nodes or in cases of rapid radiations. When choosing a probe set to start a new project, it will be important to decide upfront whether integration with previous data sets is an important factor and to choose whichever probe set was used before. Both probe kits are manufactured by the same company and have identical laboratory protocols, although Compositae1061 is slightly cheaper due to the lower number of probes per reaction. Hendriks et al. (2021) present the possibility of integrating Angiosperms353 and a custom probe set in the same hybridization reaction, which has not yet been tested in the Compositae, but is surely an exciting possibility.

We conclude that the Compositae1061 kit provides more loci, even with higher levels of paralogy, than Angiosperms353, which can be useful when working on shallower phylogenetic levels. The Angiosperms353 set yields a more even number of loci across samples that are less affected by paralogy, which can be useful when working across several lineages in the Compositae family. The outlook for phylogenomic studies in the Compositae is promising, especially if researchers across the globe are able to combine genomic data to address the evolutionary history of this large and complex group of flowering plants.

ACKNOWLEDGMENTS

C.M.S. would like to thank Matt Johnson (Texas Tech University) for initial discussions on the subject and Vicki Funk for having the idea of the Compositae1061 probe set and bringing synantherologists across the globe together to work on it. We thank Linda Watson (Oklahoma State University), Katy Jones (Botanischer Garten und Botanisches Museum Berlin), and Ramhari Thapa (University of Memphis) for samples, laboratory work, and discussions about the Compositae probe set, and Grace E. Brewer and Niroshini Epitawalage (Royal Botanical Gardens, Kew) for support in laboratory work. We thank the High Performance Computing team of the University of Memphis for computational resources and assistance. Funding came from various sources, including the National Science Foundation Division of Environmental Biology (DEB-1745197), and grants from the Calleva Foundation and the Sackler Trust to the Plant and Fungal Trees of Life (PAFTOL) project at the Royal Botanic Gardens, Kew.

AUTHOR CONTRIBUTIONS

C.M.S. and J.R.M. had the initial idea for this work. C.M.S., J.R.M., O.H., L. Palazzesi, and J.P. planned this work. J.R.M., O.M., I.J.L., F.F., and W.J.B. oversaw data production and provided the sequences used here. C.M.S. and J.R.M. conducted data analyses. C.M.S. wrote the initial draft of the manuscript, and J.R.M., O.H., L. Palazzesi, J.P., and L. Pokorny provided additional text. All authors read, commented on, and approved the manuscript and the subsequent review.

DATA AVAILABILITY

All gene trees, species trees, reference files, and code used in this article are available from https://github.com/carol-siniscalchi/Comp1061-Angio353. Raw sequence files are available at the National Center for Biotechnology Information Sequence Read Archive under BioProject PRJNA540287 and from http://paftol.org.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

APPENDIX S1. Summary of hits of individual Angiosperms353 loci sequences BLAST-searched against the Compositae1061 database.

APPENDIX S2. Basic statistics from each assembly obtained with the hybpiper_stats.py in HybPiper.

APPENDIX S3. Basic assembly statistics for each treatment and locus.

APPENDIX S4. Loci recovered with inverted assemblies.

APPENDIX S5. Paralogous loci.

APPENDIX S6. Loci flagged as paralogous for each sample.

APPENDIX S7. Loci assembled with the opposite reference and corresponding paralogous loci.

LITERATURE CITED

Abadi, S., D. Azouri, T. Pupko, and I. Mayrose. 2019. Model selection may not be a mandatory step for phylogeny reconstruction. *Nature Communications* 10: 934.

Bajgain, P., B. A. Richardson, J. C. Price, R. C. Cronn, and J. A. Udall. 2011. Transcriptome characterization and polymorphism detection between subspecies of big sagebrush (*Artemisia tridentata*). *BMC Genomics* 12: 370.

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

Barker, M. S., N. C. Kane, M. Matvienko, A. Kozik, R. W. Michelmore, S. J. Knapp, and L. H. Rieseberg. 2008. Multiple paleopolyploidizations during the evolution of the Asteraceae reveal parallel patterns of duplicate gene retention after millions of years. *Molecular Biology and Evolution* 25: 2445–2455.

Barker, M. S., Z. Li, T. I. Kidder, C. R. Reardon, Z. Lai, L. O. Oliveira, M. Scascitelli, and L. H. Rieseberg. 2016. Most Compositae (Asteraceae) are descendants

of a paleo-hexaploid and all share a paleotetraploid ancestor with the Calyceraceae. *American Journal of Botany* 103: 1203–1211.

Beck, J. B., M. L. Markley, M. G. Zielke, J. R. Thomas, H. J. Hale, L. D. Williams, and M. G. Johnson. 2021. Is Palmer's elm leaf goldenrod real? The Angiosperms353 kit provides within-species signal in *Solidago ulmifolia* s.l. bioRxiv 2021.01.07.425781 [Preprint]. Posted 8 January 2021 [accessed 2 April 2021]. Available from: https://doi.org/10.1101/2021.01.07.425781.

Bolger, A. M., M. Lohse, and B. Usadel. 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., J. J. Clarkson, O. Maurin, A. R. Zuntini, V. Barber, S. Bellot, N. Biggs, et al. 2019. Factors affecting targeted sequencing of 353 nuclear genes from herbarium specimens spanning the diversity of angiosperms. *Frontiers in Plant Science* 10: 1102.

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

Delseni, M., B. Han, and Y. I. Hsing. 2010. High throughput DNA sequencing: The new sequencing revolution. *Plant Science* 179: 407–422.

Dodsworth, S., L. Pokorny, M. G. Johnson, J. T. Kim, O. Maurin, N. J. Wickett, F. Forest, and W. J. Baker. 2019. Hyb-Seq for flowering plant systematics. *Trends in Plant Science* 24: 887–891.

Faircloth, B. C. 2015. PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics* 32: 786–788.

Godden, G. T., I. E. Jordon-Thaden, S. Chamala, A. A. Crowl, N. García, C. C. Germain-Aubrey, J. Michael Heaney, et al. 2012. Making next-generation sequencing work for you: Approaches and practical considerations for marker development and phylogenetics. *Plant Ecology & Diversity* 5: 427–450.

Hendriks, K., T. Mandáková, N. M. Hay, E. Ly, A. H. van Huysduynen, R. Tamrakar, S. K. Thomas, et al. 2021. The best of both worlds: Combining lineage-specific and universal bait sets in target-enrichment hybridization reactions. *Applications in Plant Sciences* 9(7): e11438.

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

Herrando-Moraira, S., J. A. Calleja, M. Galbany-Casals, N. Garcia-Jacas, J.-Q. Liu, J. López-Alvarado, J. López-Pujol, et al. 2019. Nuclear and plastid DNA phylogeny of tribe Cardueae (Compositae) with Hyb-Seq data: A new subtribal classification and a temporal diversification framework. *Molecular Phylogenetics and Evolution* 137: 313–332.

Huang, C.-H., C. Zhang, M. Liu, Y. Hu, T. Gao, J. Qi, and H. Ma. 2016. Multiple polyploidization events across Asteraceae with two nested events in the early history revealed by nuclear phylogenomics. *Molecular Biology and Evolution* 33: 2820–2835.

Jantzen, J. R., P. Amarasinghe, R. A. Folk, M. Reginato, F. A. Michelangeli, D. E. Soltis, N. Cellinese, and P. S. Soltis. 2020. A two-tier bioinformatic pipeline to develop probes for target capture of nuclear loci with applications in Melastomataceae. *Applications in Plant Sciences* 8: e11345.

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

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

Jones, K. E., T. Fér, R. E. Schmickl, R. B. Dikow, V. A. Funk, S. Herrando-Moraira, P. R. Johnston, et al. 2019. An empirical assessment of a single family-wide hybrid capture locus set at multiple evolutionary timescales in Asteraceae. *Applications in Plant Sciences* 7: e11295.

Katoh, K., and D. M. Standley. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* 30(4): 772–780.

Larridon, I., T. Villaverde, A. R. Zuntini, L. Pokorny, G. E. Brewer, N. Epitawalage, I. Fairlie, et al. 2020. Tackling rapid radiations with targeted sequencing. *Frontiers in Plant Science* 10: 1665.

Leebens-Mack, J. H., M. S. Barker, E. J. Carpenter, M. K. Deyholos, M. A. Gitzendanner, S. W. Graham, I. Grosse, et al. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574: 679–685.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754–1760.

Li, Z., and M. S. Barker. 2020. Inferring putative ancient whole-genome duplications in the 1000 Plants (1KP) initiative: Access to gene family phylogenies and age distributions. *GigaScience* 9(2): giaa004.

Lichter-Marck, I. H., W. A. Freyman, C. M. Siniscalchi, J. R. Mandel, A. Castro-Castro, G. Johnson, and B. G. Baldwin. 2020. Phylogenomics of Perityleae (Compositae) provides new insights into morphological and chromosomal evolution of the rock daisies. *Journal of Systematics and Evolution* 58: 853–880.

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 Sciences* 2: 1300085.

Mandel, J. R., R. B. Dikow, and V. A. Funk. 2015. Using phylogenomics to resolve mega-families: An example from Compositae. *Journal of Systematics and Evolution* 53: 391–402.

Mandel, J. R., M. S. Barker, R. J. Bayer, R. B. Dikow, T.-G. Gao, K. E. Jones, S. Keeley, et al. 2017. The Compositae Tree of Life in the age of phylogenomics. *Journal of Systematics and Evolution* 55: 405–410.

Mandel, J. R., R. B. Dikow, C. M. Siniscalchi, R. Thapa, L. E. Watson, and V. A. Funk. 2019. A fully resolved backbone phylogeny reveals numerous dispersals and explosive diversifications throughout the history of Asteraceae. *Proceedings of the National Academy of Sciences, USA* 116: 14083–14088.

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

McLay, T. G. B., J. L. Birch, B. F. Gunn, W. Ning, J. A. Tate, L. Nauheimer, E. M. Joyce, et al. 2021. New targets acquired: Improving locus recovery from the Angiosperms353 probe set. *Applications in Plant Sciences* 9(7): e11420.

Nora, S., S. Castro, J. Loureiro, A. C. Gonçalves, H. Oliveira, M. Castro, C. Santos, and P. Silveira. 2013. Flow cytometric and karyological analyses of *Calendula* species from Iberian Peninsula. *Plant Systematics and Evolution* 299: 853–864.

Ogutcen, E., C. Christe, K. Nishii, N. Salamin, M. Möller, and M. Perret. 2021. Phylogenomics of Gesneriaceae using targeted capture of nuclear genes. *Molecular Phylogenetics and Evolution* 157: 107068.

Panero, J. L., and B. S. Crozier. 2016. Macroevolutionary dynamics in the early diversification of Asteraceae. *Molecular Phylogenetics and Evolution* 99: 116–132.

Pelser, P. B., and L. Watson. 2009. Introduction to Asteroideae. *In* V. A. Funk, A. Susanna, T. F. Stussey, and R. J. Bayer [eds.], Systematics, evolution, and biogeography of Compositae, 495–502. International Association for Plant Taxonomy (IAPT), Vienna, Austria.

Plume, O. 2015. Hybridization, genome duplication, and chemical diversification in the evolution of *Calendula* L. (Compositae). Ph.D. dissertation, Cornell University, Ithaca, New York, USA.

Semple, J. C., and K. Watanabe. 2009. A review of chromosome numbers in Asteraceae with hypotheses on chromosomal base number evolution. *In* V. A. Funk, A. Susanna, T. F. Stussey, and R. J. Bayer [eds.], Systematics, evolution, and biogeography of Compositae, 61–72. International Association for Plant Taxonomy (IAPT), Vienna, Austria.

Shah, T., J. Schneider, O. Maurin, W. J. Baker, F. Forest, V. Savolainen, I. Darbyshire, and I. Larridon. 2021. Joining forces in Ochnaceae phylogenomics: A tale of two targeted sequencing probe kits. *American Journal of Botany* 108(7): https://doi.org/10.1002/ajb2.1682

Shee, Z. Q., D. G. Frodin, R. Cámara-Leret, and L. Pokorny. 2020. Reconstructing the complex evolutionary history of the Papuasian *Schefflera* radiation through herbariomics. *Frontiers in Plant Science* 11: 258.

Siniscalchi, C. M., B. F. P. Loeuille, V. A. Funk, J. R. Mandel, and J. R. Pirani. 2019a. Phylogenomics yields new insight into relationships within Vernonieae (Asteraceae). *Frontiers in Plant Science* 10: 1224.

Siniscalchi, C. M., B. Loeuille, J. R. Pirani, and J. R. Mandel. 2019b. Using genomic data to develop SSR markers for species of *Chresta* (Vernonieae, Asteraceae) from the Caatinga. *Brazilian Journal of Botany* 42: 661–669.

Slater, G., and E. Birney. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6: 31.

Slimp, M., L. D. Williams, H. Hale, and M. G. Johnson. 2021. On the potential of Angiosperms353 for population genomic studies. *Applications in Plant Sciences* 9(7): e11419.

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

Smith, S. A., N. Walker-Hale, J. F. Walker, and J. W. Brown. 2020. Phylogenetic conflicts, combinability, and deep phylogenomics in plants. *Systematic Biology* 69: 579–592.

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

Straub, S. C. K., M. Parks, K. Weitemier, M. Fishbein, R. C. Cronn, and A. Liston. 2012. Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. *American Journal of Botany* 99: 349–364.

Susanna, A., B. G. Baldwin, R. D. Bayer, J. M. Bonifacino, N. Garcia-Jacas, S. C. Keeley, J. R. Mandel, et al. 2020. The classification of the Compositae: A tribute to Vicki Ann Funk (1947–2019). *Taxon* 69(4): 807–814.

Thapa, R., R. J. Bayer, and J. R. Mandel. 2019. Development and characterization of microsatellite markers for *Antennaria corymbosa* (Asteraceae) and close relatives. *Applications in Plant Sciences* 7: e11268.

Thapa, R., R. J. Bayer, and J. R. Mandel. 2020. Phylogenomics resolves the relationships within *Antennaria* (Asteraceae, Gnaphalieae) and yields new insights into its morphological character evolution and biogeography. *Systematic Botany* 45: 387–402.

Ufimov, R., V. Zeisek, S. Píšová, W. J. Baker, T. Fér, M. van Loo, C. Dobeš, and R. Schmickl. 2021. Relative performance of customized and universal probe sets in target enrichment: A case study in subtribe Malinae. *Applications in Plant Sciences* 9(7): e11442.

Van Andel, T., M. A. Veltman, A. Bertin, H. Maat, T. Polime, D. Hille Ris Lambers, J. Tjoe Awie, et al. 2020. Hidden rice diversity in the Guiana. *Frontiers in Plant Science* 10: 1161.

Van de Peer, Y., E. Mizrachi, and K. Marchal. 2017. The evolutionary significance of polyploidy. *Nature Reviews Genetics* 18: 411–424.

Watson, L., C. M. Siniscalchi, and J. R. Mandel. 2020. Phylogenomics of the hyperdiverse daisy tribes: Anthemideae, Astereae, Calenduleae, Gnaphalieae, and Senecioneae (Asteraceae). *Journal of Systematics and Evolution* 58: 841–852.

Wendel, J. F. 2015. The wondrous cycles of polyploidy in plants. *American Journal of Botany* 102: 1753–1756.

Zhang, C., M. Rabiee, E. Sayyari, and S. Mirarab. 2018. ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* 8: 153.