THESIS

PARALOGY OR REALITY? EXPLORING GENE ASSEMBLY ERRORS IN A TARGET ENRICHMENT DATASET

Submitted by

Austin Rosén

Department of Biology

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Summer 2022

Master's Committee:

Advisor: Mark P. Simmons

Jennifer Ackerfield
Christopher Richards
Jane Stewart

ABSTRACT

PARALOGY OR REALITY? EXPLORING GENE ASSEMBLY ERRORS IN A TARGET ENRICHMENT DATASET

De novo gene assembly of short read data is inherently difficult – similar to the process of assembling a jigsaw puzzle. I describe three errors that occurred with the assembly of target enrichment data in the genus *Cirsium* (Asteraceae): inconsistent contig selection, artificial recombination, and inconsistent intron determination leading to over-alignment of non-homologous nucleotides. These errors occurred in 39% of loci in the dataset and were often a by-product of undetected paralogs: assembled loci that likely contained paralogous or homoeologous sequences but did not trigger default paralog warnings by the assembly program, HybPiper. Default HybPiper thresholds for identifying paralogy during the assembly process were insufficient to filter such loci. A custom target file was created in which putative paralogs were separated into independent loci. The custom target file was successful in reducing, but not eliminating, assembly errors in the dataset. A final iteration of quality control was performed to create a dataset largely free of assembly errors. However, phylogenetic inferences applied to this final cleansed dataset were unable to resolve the taxonomic relationships between the sampled specimens. Rather, these results affirm that *Cirsium* is a taxonomically problematic genus and may require population-level genetic data or integrative taxonomy approaches to delimit species boundaries.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my advisor Mark Simmons for his constant and consistent guidance and support over the past three years. I would also like to thank my committee members Jennifer Ackerfield, Chris Richards, and Jane Stewart, for their guidance and assistance in completing this project. In particular, Jennifer Ackerfield provided me with much needed support and mentorship during the unavoidable rough patches that occur during one's graduate school experience. I would like to give thanks to Wendy Hodgson of the Desert Botanic Gardens for inviting me along on two unforgettable collecting trips in the American southwest, and Carolina Siniscalchi and Pat Reeves for valuable direction and insight regarding my sequence data. I would also like to thank Teia Schweizer, Luke Tembrock, Gus Waneka, and the Dan Sloan Lab for providing essential lab equipment, Alaina Krakowiak for the guidance she provided for performing and improving DNA isolations, and my parents Lee Rosén and Priscilla Virant for their encouragement, support, and home-cooked meals. I thank all the herbaria that provided loans for this project: ASC, ASU, BRY, DES, OBI, RENO, RSA, and UCR. Funding for this work was provided in part by the Margaret Williams Research Grant from the Nevada Native Plant Society, the Bill Dahl Graduate Student Research Award from the Botanical Society of America, and Colorado State University's Harold David Harrington Graduate Research Fellowship and Stavros Family Fund Scholarship. Finally, and most importantly, I would like to thank my partner Anna West for her unyielding support and dedication to my success, despite the many sacrifices required over the past three years to make this possible.

TABLE OF CONTENTS

INTRODUCTION

High-throughput sequencing methods enable researchers to incorporate genome-wide data into phylogenetic studies. However, the short sequence reads produced by such techniques typically require substantial computational effort to assemble. *De novo* gene assembly of short read data is inherently difficult – analogous to the process of assembling a jigsaw puzzle (Schatz et al., 2010). These challenges are often exacerbated in plant taxa for several biological and computational reasons (Schatz et al., 2012). Plant genomes can be quite complex, with genome duplication events leading to large gene families consisting of nearly identical sequences, such that *de novo* gene assembly can be more challenging in plants compared to many other taxa (Schatz et al, 2012).

Sequencing degraded DNA from herbarium specimens can result in differential amplification success, and lower quality reads of shorter length leading to recovery of shorter, lower quality contigs (McKain et al., 2018). Differential amplification success may result in a single sequence being recovered for each specimen for a gene family, potentially resulting in the comparison of non-orthologous gene copies. Orthologs are homologous gene copies that were inherited from a common ancestor and diverged from one another due to speciation events, usually retaining their original function. By contrast, paralogs are homologous gene copies arising from duplication events, and often acquire a new function altogether. Determining evolutionary relationships between species requires comparison of orthologs. The comparison of paralogs has potential to produce misleading phylogenetic results (Fitch, 1970; Doyle, 1992). For gene families with low sequence divergence among paralogs, *de novo* assembly of short reads can result in assembly errors in which the output sequence does not represent any true ortholog. The introduction of such errors is likely to result in false biological conclusions from downstream analyses (Schatz et al., 2012).

Case Study of *Cirsium mohavense*

I present a case study of gene assembly errors, their effect on downstream analysis, and a potential solution to such errors. I describe three errors that occurred with the assembly of target enrichment data in the flowering plant genus *Cirsium* (Asteraceae), commonly known as thistles. A target enrichment approach was chosen due to its ability to recover hundreds of pre-determined loci, the ability to combine data across studies that targeted the same loci, its ability to capture flanking intron regions that are helpful for resolving relationships at shallow taxonomic levels, and its resilience to degraded DNA typical of herbarium specimens (McKain et al., 2018; Herrando-Moraira et al., 2018). This sequencing approach was expected to be appropriate for this study, in which most of the available leaf tissue came from herbarium specimens, and specimens processed from different batches of library preparation and sequencing that were combined in the final dataset.

Numerous bioinformatics pipelines have been developed to extract and assemble sequence reads acquired from target enrichment methods, such as HybPiper (Johnson et al. 2016), Phyluce (Faircloth, 2016), HybPhyloMaker (Fér & Schmickl, 2018), and Assexon (Yuan et al., 2019). Multiple studies have been published that make direct comparisons between these assembly pipelines, and despite the differences in their assembly methods, the results they produce are very similar (Fér & Schmickl, 2018; Herrando-Moraira et al., 2018; Yuan et al., 2019). HybPiper was chosen for this study because it was specifically developed for target enrichment protocols in which the probes were designed from exon sequences, such as the probes used in this study (Mandel et al., 2014; Johnson et al., 2016). In contrast, Phyluce was specifically designed for probes designed using ultra conserved elements (Faircloth, 2016), which are more commonly used for target enrichment approaches applied to animal taxa (McKain et al., 2018). Throughout the investigation of gene assembly errors, I found that HybPiper outputs are easy to work with, and do not require any advanced bioinformatics skills – it is quite easy to examine the output of programs utilized by HybPiper and investigate various steps of the assembly process.

Focal Taxa

Cirsium has long been recognized as a taxonomically problematic genus. Indistinct morphological differentiation, frequent hybridization, incipient speciation, and sampling inadequacies have made the delimitation of species boundaries within much of the genus problematic (Welsh, 1982, 1983; Kelch & Baldwin, 2003; Keil, 2006). Often large and armed with spines, field collections of *Cirsium* can be difficult to make, leading to under-collection of the genus, and herbarium specimens that often contain only a small portion of the original plant (Welsh, 1982). Consequently, morphological variation within species is either misrepresented or underrepresented in herbaria and therefore is poorly understood (Welsh, 1982; Keil, 2006).

One species complex within *Cirsium* that has had taxonomic instability consists of *C. mohavense* (Greene) Petr., *C. virginense* Welsh, and "*C. walapaiorum*," which are found in arid seeps and springs throughout the Mojave Desert and southern Great Basin. These species have shared morphological traits of dense gray tomentum throughout, involucral bracts with a glutinous dorsal ridge, strongly decurrent leaf bases, and white to pink or lavender corollas (Keil, 2006). Wetland obligates, these plants are rhizomatous perennials or biennials, and can grow to four meters. *Cirsium virginense*, which occurs in the Virgin River Basin, is treated as a synonym of *C. mohavense* sensu Keil (2006) in *Flora of North America*, but it is still considered to be a rare species of conservation concern in both Nevada and Utah (Utah Native Plant Society, 2003; Nevada Natural Heritage Program, 2020). "*Cirsium walapaiorum"* is a proposed new taxon within the Grand Canyon (Hodgson & Rink, 2018) based on morphological evidence and geographic distribution, but has not been formally described. Given the taxonomic disagreement within the group, and currently recognized taxonomic difficulties with *Cirsium*, I performed a study incorporating molecular data to determine the number of independent evolutionary lineages (i.e., species) present under the general lineage concept (de Quieroz, 1998).

This is a case study of gene assembly errors occurring with target capture data using the HybPiper bioinformatics pipeline. The main objectives of this study are (i) to determine the underlying causes of the gene assembly errors, (ii) to test the effectiveness of a solution at resolving the observed gene assembly errors, and (iii) determine if any phylogenetic inference approach applied was useful for resolving the relationships among the specimens sampled.

MATERIALS AND METHODS

Sampling

Seventy specimens of *Cirsium mohavense* (sensu lato) were sampled, as well as six outgroup specimens of North American *Cirsium*. Specimens sampled included herbarium collections and silica-dried leaf material from plants collected in 2019 (Appendix A). Library preparation and sequencing for these 76 total specimens were completed in two batches with slightly different methods. Twelve specimens were sequenced in the first batch, and the remaining 64 specimens were sequenced in the second batch. Post-sequencing bioinformatics methods were identical for all specimens used in the analysis.

DNA Extraction

For the first batch of specimens, leaf tissue was ground with a Mixer Mill MM 301 (Retsch, Haan, Germany) and genomic DNA was extracted with a DNeasy Plant Mini Kit (Qiagen, Valencia, CA, USA) according to the manufacturer's instructions. For the second batch, genomic DNA was extracted from leaves using a modified version of the protocol described by Alexander et al. (2006). Samples were ground at 30 hertz for 60 seconds with a TissueLyser II (Qiagen, Hilden, Germany), rather than the grinding method described in the paper. DNA concentration from all specimens was quantified using a Qubit 3.0 Fluorometer (Thermo Scientific, Waltham, MA, USA) and the size distribution measured by using a 1% agarose gel.

Library Preparation, Target Enrichment, and Sequencing

For the first batch of samples, genomic DNA was fragmented into ~400 bp segments using a Q800R2 Sonicator (Qsonica LLC, Newton, CT, USA) and confirmed with a 1% agarose gel. Library preparation was performed with the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, MA, USA). For the second batch, library preparation was performed with the NEBNext

Ultra II FS DNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, MA, USA), with enzymatic fragmentation times optimized to produce fragments between ~250-400 bp long and confirmed with a 1% agarose gel. NEBNext Multiplex Oligos for Illumina were used to barcode samples, and AMPureXP magnetic beads (Beckman Coulter, La Brea, CA, USA) were used for size selection. The concentration of each sample in the library was quantified using a Qubit 3.0 fluorometer and the size distribution measured by using a 1% agarose gel. AMPureXP magnetic beads were used for size selection, and individual libraries were quantified using a Qubit 3.0 fluorometer. Samples from the first batch were pooled into groups of four with ~500 ng of the total library from each sample. Samples from the second batch were pooled into groups of 22 or 23 samples with ~100 ng of total library from each sample. Library pools were evaporated using a vacuum centrifuge and immediately rehydrated with 7 µl of dH_2O.

Target capture was performed using myBaits Expert Predesigned Panel (Arbor Biosciences, Ann Arbor, MI, USA) COS Compositae 1Kv1 (also known as the Composita1061 probe set; Mandel et al., 2014), allowing hybridization reactions to occur for 40 hours at 65° C. KAPA HiFi (2´) HotStart ReadyMix PCR Kit (Roche, Basel, Switzerland) was used for post-capture PCR amplification for the first batch of samples. For the second batch, post-capture PCR amplification was performed used NEBNext Ultra II Q5 Master Mix with 20 cycles. AMPureXP beads were used for a final clean-up of the PCR products and final enriched library concentrations were measured using a Qubit 3.0 fluorometer. These final pools were sequenced at Psomagen (Rockville, MD, USA) on an Illumina HiSeqX (2 × 150 bp paired-end reads).

Raw Data Processing and Target Enrichment Extraction

Raw sequence reads were demultiplexed by the sequencing facility and checked for basic read quality using FastQC v.0.11.9 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Raw FASTQ reads were trimmed and adaptor content removed using Trimmomatic v.0.39 (Bolger et al., 2014), with options set to: (1) remove leading low quality or N bases with a Phred-scaled quality score below 20, (2) remove trailing low quality bases below quality 20, (3) scan the read with a 5-bp sliding window, cutting when the average quality within each window drops below 20, and (4) exclude reads < 36 bases long. The processed paired reads were then assembled using HybPiper v.1.3.1 (Johnson et al., 2016) with the BWA (Burrows-Wheeler aligner) option to map reads to the targets contained in the target file. The target file provided to HybPiper for the initial data processing was the source expressed sequence tags (ESTs) for the probes used in the COS Compositae target capture design and included one to three orthologous sequences for the 1061 loci targeted in the myBaits Compositae kit (Mandel et al., 2014). The coding sequences recovered for each locus were retained for downstream analyses.

Sequence Alignment and Alignment Quality-Control

The multi-fasta files produced by HybPiper for each locus were aligned individually using MAFFT v.7.48 (Katoh and Standley, 2013) with the L-INS-I algorithm, which is the most accurate MAFFT algorithm for non-rDNA loci, and maximum iterations were set to 1,000. Following Simmons et al. (2022) we used trimAl v.1.4 (Capella-Gutiérrez et al., 2009) to implement the following five quality control steps: (1) discard inconsistent alignment positions between head-and-tails alignments (Landan and Graur, 2007), (2) discard alignment positions that contain gaps in greater than 90% of the sequences, (3) discard hypervariable alignment positions based on a nucleotide-similarity threshold of 0.001, and (4) discard entire sequences when less than 50% of their nucleotide characters have an overlap score of 0.5. Fifth, TAPER v.0.1.6 (Zhang et al., 2021) was used to mask regions of individual sequences that were divergent outliers in comparison to the rest of the sequences in the alignment by calling the script correction_multi.jl with the options -c 1 and -m N. Finally, all genes flagged by HybPiper as containing potential paralogs in any specimen were excluded from the dataset before phylogenetic analysis to help avoid potential comparison of non-orthologous sequences.

Problematic Observations

Visual inspection of multiple-sequence-alignment files (after alignment quality-control steps) in Geneious Prime v.2021 (Kearse et al., 2012) showed patterns among numerous genes (e.g., Figure 1A) that were almost identical to patterns seen in genes flagged as potential paralogs (e.g., Figure 1B) by HybPiper. Figures 1A and 1B show an exemplar subset of specimens. There were sets of sequences that were conserved relative to each other but highly divergent relative to one or more other sets of sequences. Undiagnosed paralogy can compromise phylogenetic inference, by conflating serial and taxic homology (Doyle, 1992). Therefore, the presence of paralog-like patterns in unflagged genes warranted investigation into whether the dataset included unflagged paralogs.

HybPiper Assembly Methods

HybPiper uses SPAdes (Bankevich et al., 2012) to conduct *de novo* assembly of trimmed reads into contiguous sequences using known protein coding DNA sequences as a reference. Multiple contiguous sequences for a specimen may be assembled for the same target locus for several reasons, including presence of paralogs, alleles from heterozygotes, incomplete intron sequencing, and reads with low sequence identity to the reference (Johnson et al., 2016). HybPiper then uses Exonerate (Slater and Birney, 2005) to extract regions from the assembled contiguous sequences that represent the coding sequence for each target gene (i.e., exon contig). By default, an exon contig is only considered for extraction if it passes a default threshold of 60% identity to the reference sequence. If only a single exon contig is extracted by Exonerate, that sequence is output by HybPiper as the coding sequence for that gene. Alternatively, If Exonerate extracts multiple exon contigs that pass the 60% identity threshold, HybPiper produces a single output sequence based on the following three criteria.

First, if the contigs are non-overlapping in alignment to the reference, or overlapping by ≤ 20 bp, HybPiper joins them together to form an “exon-supercontig.” Hereafter the term “exon-supercontig” will be used to refer to such cases, to prevent confusion with supercontigs comprised of both exon and flanking regions, which will be referred to hereafter as “intron-supercontigs.” An additional Exonerate search is then performed by HybPiper to identify the correct junction between exon and intron regions. Second, if one contig completely subsumes the range (when aligned to the reference) of all other contigs extracted by Exonerate, the longest contig is chosen, regardless of its identity to the reference or depth of coverage. Third, if two or more long contigs are extracted that span at least 85% of the reference, a paralog warning is triggered by HybPiper. A single long contig is then selected for output by considering depth of coverage and identity to the reference sequence. A contig is selected if it has at least 10× greater coverage than all other competing contigs. If no contig meets this criterion, the contig with the greatest percent identity to the reference is chosen (Johnson et al., 2016). Users can then retrieve the final coding sequence (contig or exon-supercontig), intron sequence, or intron-supercontig sequence output by HybPiper for each specimen. For this study, only coding sequences were retained for downstream analyses, while intron sequences and intron-supercontig sequences were used to aid visual investigation of gene assemblies.

Assembly Investigation

To check for undetected paralogs, a custom bash script was used to retrieve all exon contigs assembled by Exonerate for each specimen for each gene, as well as the final exon sequence (contig/exon-supercontig) output by HybPiper. These sequences were aligned in MAFFT and manually examined to check for assembly errors. Intron-supercontigs were not investigated in this manner, because they are produced by HybPiper after it chooses the final output contig/exon-supercontig. Therefore, comparisons across intron-supercontigs would not help identify causes of assembly errors resulting from multiple contig recovery. For many genes, Exonerate recovered multiple contigs for two or more specimens. Sometimes these contigs overlapped extensively but were not flagged by HybPiper as potential paralogs, because they did not pass the paralog warning threshold (85% of the target length). Investigation of the contigs recovered by Exonerate for these genes showed a pattern of competing contigs (Figure 2A), similar to the pattern of competing contigs observed for genes that triggered HybPiper's paralog warning (Figure 2B). I therefore considered genes that showed patterns indicative of paralogy but not flagged by HybPiper to be undetected paralogs. Gene assemblies that included undetected paralogs were likely the result of the following three scenarios.

Inconsistent Contig Selection

In the first scenario, the exon contig output by HybPiper was inconsistent across specimens, either despite similar recovery of competing contigs or because of differential recovery of competing contigs. For the specimens shown in Figure 2A, when multiple contigs were recovered but that gene was not flagged as a paralog, the output contigs selected by HybPiper's algorithm followed the description in the program documentation for paralogous genes. When the longest of the competing exon contigs completely subsumed the range of all other contigs for that specimen, that longest contig was chosen as the output sequence. This scenario can be seen for specimens 369 and 372 in Figure 2A, in which the outputs are contig A and contig C, respectively. Alternatively, if no contig completely subsumed the range of all other contigs, length was discarded by HybPiper as a criterion, and if no contig had at least 10× greater coverage than all other competing contigs, the contig with the greatest percent identity to the reference was chosen. This can be seen for specimen 301 in Figure 2A, for which the output was contig B. For both paralogs and genes not flagged as paralogs by HybPiper, differences in contig recovery, contig length, and depth of coverage resulted in inconsistent selection of *in vivo* sequences by HybPiper.

Artificial Recombination

In the second scenario, artificial recombination was observed such that the output sequence likely does not represent any *in vivo* ortholog. This artificial recombination occurred only when the output sequence for a specimen was an exon-supercontig. Many cases were observed in which the contigs that were joined together to form the output sequence were inconsistent across specimens. In some cases, this inconsistency was caused by incomplete recovery of competing contigs, and in other cases, occurred despite recovery of almost identical contigs. Both sets of cases resulted in artificial recombination wherein contigs were joined together in various ways, with the number of potential combinations increasing with the number of contigs. The latter case is illustrated in Figure 3, which shows that the 5' end of the output sequence for specimens 179 and 341 originate from the same contig, while the 3' end of output sequence originate from different contigs. Furthermore, despite recovery of almost identical competing contigs to specimens 179 and 341, artificial recombination was not observed for specimen 401. Inconsistent contig selection and artificial recombination often co-occurred in problematic cases of undetected paralogy (Figure 4; inconsistent contig selection for specimens 301 and 369; artificial recombination for specimens 84 and 340), and even occurred between extensively overlapping contigs (Figures 3 and 4). The recombination breakpoint was inconsistent among recombinant outputs in Figure 4 (specimens 84 and 340), in contrast to the consistent breakpoint observed in Figure 3.

Over-alignment and Inconsistent Intron Determination

In the third, most complex scenario, non-overlapping contigs were recovered for a specimen, but the final exon-supercontig output by HybPiper included additional sequence data between the contigs (Figure 5A). In the multiple-sequence-alignment, these regions were highly divergent, with many more variable positions than the rest of the alignment (Figure 5A). To determine where these additional sequences originated, the final intron-supercontig sequence, intron sequence, assembled contigs, and final exon-supercontig were retrieved from HybPiper and aligned using MAFFT (Figure 5B). This investigation showed that the hypervariable region in question came from the intron-supercontig sequence minus the intron sequence (Figure 5B), but was not included in either of the assembled exon contigs for either specimen shown in Figure 5A. By definition this alignment region should represent the exon sequence, and should be unproblematic, so its extreme divergence in the multiple-sequence-alignment was concerning. Presumably, this hypervariable sequence region was incorporated into the output sequence when Exonerate performed its additional search to identify the correct junction between exon and intron regions after forming the exon-supercontig. However, my methods of investigation were unable to confirm the reason why the hypervariable region became incorporated into the output sequence. In other specimens, such as specimen 358 shown in Figure 5B, this same region was determined by HybPiper to be intron sequence and was therefore not included in the final output. It appears that HybPiper was inconsistent across specimens in its interpretation of exons versus introns. In Figure 5B, when aligned along with introns, this region did not appear highly divergent. However, when only the putative exons were used, the region became mis-aligned by MAFFT and appeared hypervariable because of over-alignment with a non-homologous region (Figure 5C). Therefore, the over-alignment observed in this original dataset was typically due to inconsistent intron determination by HybPiper followed by over-alignment of non-homologous nucleotides by MAFFT.

Proposed Solution

Following a proposed solution for dealing with paralogs described in HybPiper's documentation (https://github.com/mossmatters/HybPiper/wiki/Paralogs), a custom target file was created for our dataset to treat putative paralogs as separate loci. Putative paralogs include those flagged by HybPiper as well as those that appeared to contain undetected paralogy during the gene investigation--even when it did not lead to the assembly errors described above. Ideally, each of these putative paralogs would align to only the reference target representing that paralog. The result should be a final gene assembly in which only a single contig is recovered for each specimen, thereby avoiding comparison of non-orthologous sequences and potential artificial recombination. This approach should increase the number of distinct loci that can be sampled for phylogenetic inference, while at the same time, reducing gene assembly errors.

Solution Test

The utility of this solution in preventing artificial recombination and inconsistent contig selection was tested for an exemplar gene. Gene At1g01050 was chosen because it contained both inconsistent contig selection by HybPiper and artificial recombination (Figure 4), and for which at least one exon contig was recovered for all 76 specimens. The original target sequence for this gene was replaced with three new target loci, each being the longest (or tied for longest) representative of one of the three competing exon contigs recovered across all specimens for that gene (Figure 6A). Gene assembly was then performed in HybPiper using the new targets.

Final At1g01050 alignments viewed after all alignment quality control steps no longer showed patterns indicative of inconsistent contig selection and/or artificial recombination, and no hypervariable regions (Figure 6B). Therefore, this solution was successful in parsing out undetected paralogs from the test gene and preventing the inconsistent contig selection and artificial recombination that occurred with the original target file.

Creation of Custom Target File

The custom target file was created using the following decision-rules (Figure 7). These decision-rules were applied to each individual gene alignment containing all exon contigs recovered for all specimens, which were created during the assembly investigation described here. Intron-supercontigs were not applicable because flanking regions are added in a step that occurs after potential gene assembly errors have already been introduced.

If only a single contig was recovered for each specimen, no changes were made to the target file for that gene. However, if more than one contig was recovered for any sampled specimen, three other criteria were assessed. First, if all contigs recovered for individual specimens were non-overlapping or overlap by ≤ 20 bp, no changes were made to the target file for that gene (Figure 7). This decision was based on my observation that inconsistent contig selection and artificial recombination did not occur when all contigs were minimally or non-overlapping.

Second, if there was > 20 bp overlap between contigs within a sampled specimen, but all overlapping contigs for that specimen had a pairwise identity ≥ 90%, no changes were made to the target file for that gene (Figure 7). In these cases, the overlapping contigs were similar enough that if inconsistent contig selection or artificial recombination were to occur, I did not expect them to severely compromise phylogenetic analyses.

Third, if there was > 20 bp overlap between the contigs recovered for a specimen and the competing contigs had a pairwise identity of less than 90%, but the number of recovered contigs for any specimen was greater than six, that gene was removed from the target file (Figure 7). This decision was based on my expectation that these genes had high potential for inconsistent contig selection among sampled specimens and/or artificial recombination during assembly.

Alternatively, if there was > 20 bp overlap between contigs within a specimen and these competing contigs had a pairwise identity < 90%, but the number of recovered contigs did not exceed six for any specimen, the target file was amended for that gene. In these cases, sequences for that gene initially included in the target file were replaced by the longest recovered sequence for each competing contig recovered across all specimens (Figure 6). To maintain consistency with the COS Compositae target file, contigs shorter than 60 bp were not included as new targets. However, some short contigs (60 – 150 bp) were included in the target file, not because of potential phylogenetic signal within these genomic regions, but as a tool to help prevent artificial recombination during assembly of other targets for that gene. All target enrichment extraction, alignment and quality-control methods described above were executed with the custom target file, which replaced the original target file.

Phylogenetic Analysis

Phylogenetic analysis was conducted for the datasets generated with each target file using both concatenation and coalescent approaches. For the concatenation analysis, individually aligned loci were combined into a single super-matrix using AMAS v.1.0 (Borowiec, 2016). Phylogenetic trees were estimated with both a parametric approach using maximum likelihood (ML) and a non-parametric approach using parsimony. To infer trees by ML, I used IQTREE v.2.1.3 (Minh et al., 2020) with the option -m MFP to determine the best fitting model for the data, and ultrafast bootstrapping (Hoang et al., 2018) with 1,000 pseudoreplicates. To infer trees by parsimony I used TNT v.1.5 (Goloboff et al., 2008) with 1,000 tree-bisection-reconnection (TBR) search replicates with up to 50 trees retained per search. I retained the strict consensus of all identified most parsimonious trees and calculated bootstrap support using 1,000 pseudoreplicates, performing 100 TBR searches in each pseudoreplicate, and retaining up to 50 trees per search.

For the coalescent approach, separate strict-consensus gene trees were estimated by parsimony for each locus using TNT, concatenated into a single file, and used to estimate a non-parametric species tree with ASTRAL v.5.7.8 (Zhang et al., 2018). Analysis of individually aligned loci showed very little sequence divergence across specimens (typically >90% sequence identity), so I chose not to apply ML for individual gene trees to avoid arbitrary resolution (Simmons and Gatesy, 2021). All phylogenetic tree analyses were rooted using *Cirsium occidentale* (specimen 78).

Assessment of Custom Target File

The following three measures were used to assess the utility of the custom target file at resolving gene assembly errors after its application. First, the number of new targets that were flagged by HybPiper's default paralog warning was recorded. Second, the number of new targets that recovered a single exon contig across all recovered specimens was recorded. Third, a test for recombination was performed for each final gene assembly by calculating the pairwise-homoplasy-index (PHI) statistic using

PhiPack, with a significance level of 0.05 (Bruen 2005, Bruen et al., 2006). To correct for the occurrence of false positives due to multiple tests, p-values for individual genes were adjusted using the False Discovery Rate (FDR) method in R v.4.1.0 (R Core Team, 2021). Genes that triggered default paralog warnings in HybPiper were not tested for recombination because they were already excluded from the dataset during initial quality control steps.

Phylogenetic inferences were also compared between the datasets generated using each target file. The degree to which the custom target dataset reduced gene-tree discordance in the species tree topology was assessed by comparing the Normalized Quartet Score (NQS) calculated by ASTRAL. For ML concatenation analysis, the amount of observed molecular divergence was compared between the two datasets using the sum of branch lengths across the tree. For parsimony, the observed degree of homoplasy was compared using the ensemble retention index (RI; Farris, 1989). Average bootstrap support among all clades was also compared between the ML trees as well as between the parsimony trees, and average local posterior probability (LPP) values were compared between the ASTRAL species trees.

Final Iteration of Quality Control

To create a final cleansed dataset for inference of relationships among the sampled *Cirsium* specimens, one last iteration of quality control was implemented. All loci flagged as recombinant by the PHI test, and all remaining loci with assembly errors were excluded from the analysis. Loci with assembly errors were identified using the same contig investigation method used for the initial discovery of assembly errors in the original dataset. All phylogenetic inferences previously described in the methods were performed for this final dataset, and gene-tree discordance, sum of branch lengths, ensemble retention index, and support values were once again examined. Finally, I explored if the data would be more appropriately examined using a phylogenetic network [NeighborNet as implemented in SplitsTree v.4.17.1 (Huson and Bryant, 2006)], which allows for reticulate relationships between specimens.

RESULTS

Sequencing and Assembly Errors

Seventy-six specimens were successfully sequenced. Using the original target file 1,030 genes were recovered, with HybPiper default paralog warnings triggered for 272 genes (Table 1). Undetected paralogy leading to inconsistent contig selection and/or artificial recombination was visually observed in 286 (28%) of the recovered genes. Over-alignment was identified in 112 genes (11%) and multiple contigs were recovered in 921 genes (89%).

Custom Target File

Alternative target loci were created for 390 of the genes (37%) in the original target file, and 73 genes (7%) were excluded from the custom target file because they contained specimens for which greater than six contigs were recovered. On average, each of the 390 altered genes were replaced with 2.6 alternative target loci, totaling 995 custom targets. The resulting custom target file comprised 1,593 total targets, including the 995 custom targets not present in the original target file (Table 1). The mean size of target sequences in the custom target file was ~74 bp (18%) shorter than the original target file (Table 1).

Using the custom target file, HybPiper recovered sequence data for 1,564 loci, a total of 534 (52%) more target loci than the original target file (Table 1). The exon contigs recovered using the custom target file were an average of ~32 bp (10%) shorter than the exon contigs recovered using the original target file (~275 compared to ~307). The custom target file slightly reduced the total proportion of loci that triggered paralog warnings but greatly reduced the average number of paralog warnings per specimen from 121 to 41 (Table 1). A default paralog warning was triggered in HybPiper for 314 (32%) of the 995 custom targets. The average number of contigs recovered per specimen for each locus was reduced by 44% (2.5 contigs to 1.4 contigs) using the custom target file, with this reduction being greatest (59%; 3.4 contigs to 1.4 contigs) among loci flagged as potential paralogs by HybPiper. A total of 153 (15%) of the custom targets resulted in assemblies in which only a single contig was recovered for all specimens for which sequence data was recovered.

The percentage of loci identified as recombinant by the PHI test (after removal of paralogs flagged by HybPiper) was reduced from 22.2% for the dataset generated using the original target file, to 9.5% for the dataset generated using the custom target file (Table 2). Conversely, the proportion of loci that failed the requirements of the PHI test because of too little sequence divergence increased by 6.6% when using the custom target file (Table 2). Of all loci flagged as recombinant in the custom target file dataset, 57% were created from new targets, comprising 6% of all new target loci that were created.

The concatenated dataset created with the custom target file had 40% more characters and 42% fewer parsimony informative characters than the original dataset (Table 3). Overall, there was low sequence diversity in both the original and custom target datasets. Parsimony informative characters comprised only 6% of the original dataset, and only 2% of the custom target dataset. Tree lengths were reduced by 65% in maximum likelihood analysis in the custom target dataset compared to the original dataset, and average bootstrap supports increased by 9% (Figures 8 and 9; Table 3). For parsimony analysis, there was a 27% increase in the ensemble retention index and a 28% increase in average bootstrap supports from the original dataset (Figure 10) to the custom target dataset (Figure 11).

For the custom target dataset, species tree inference using ASTRAL had 17% higher average local posterior probabilities (LPP) and a 26% higher normalized quartet score (NQS) than the original dataset (Figures 12 and 13; Table 4).

Final Iteration of Quality-Control

The 38 loci identified as recombinant after using the custom target file were excluded from the final dataset. Visual examination of these alignments after all quality-control steps identified seven loci still in the dataset with assembly errors related to undetected paralogy (i.e., inconsistent contig selection and/or artificial recombination) and nine loci with over-alignment. These 16 loci were also excluded from the final dataset. After all alignment quality-control steps were implemented, the final cleansed dataset comprised 1124 loci, amounting to a total increase of 380 loci from the original target dataset (Table 3). Eighty two percent of the assembled loci recovered multiple contigs in one or more specimens, a decrease from 90% in the original dataset.

The final dataset had both fewer total characters and fewer parsimony informative characters than the custom target dataset (Table 3). Similar to the custom target dataset, parsimony informative characters comprised ~2% of the final cleansed dataset. For maximum likelihood analysis (Figure 14), the sum of branch lengths was further reduced, and average bootstrap support values further increased (Table 3). For parsimony analysis (Figure 15), the final dataset had a slightly lower ensemble retention index and slightly lower average bootstrap support values per node compared to the custom target dataset. Species-tree inference using ASTRAL (Figure 16) had no difference in average support value (LPP) per node than the custom target dataset, and only slightly lower topological discordance (i.e., higher NQS score; Table 4).

The NeighborNet network for the final cleansed dataset (Figure 17) did not show obvious clustering of any large groups of specimens that appeared highly distinct from any other groups of specimens. Rather, the general pattern of the network was short boxy (i.e., conflicting characters in a hierarchical context) internal branches in comparison to long terminal branches, similar to the trees inferred by ML methods. For example, there was obvious conflicting hierarchical signal for specimens 148, 288, and 389. While the network showed evidence of a reticulate relationship among specimens in the study, the relative scale of the network was dominated by the long terminal branches resulting from inferred autapomorphies.

DISCUSSION

Utility of the Custom Target File

Overall, use of the custom target file was successful at increasing the number of loci that were recovered from the raw sequence data (Table 1), reducing the average number of contigs recovered per locus (Table 1), and reducing the frequency of artificially recombinant loci (Table 2). Although the custom targets were an average of 18% shorter than the original targets, the length of exon contigs recovered by HybPiper were only 10% shorter using the custom target file compared to using the original target file. Reduction in the number of parsimony informative characters in the custom target dataset was the result of separation of non-orthologous sequences created by observed assembly errors. The shorter summed branch lengths on the ML tree, fewer steps on the parsimony tree, and increase in the ensemble retention index also indicate less character conflict than the original dataset. Similarly, the species-tree topology for the final dataset had lower discordance and higher average branch supports than the original dataset. Despite these apparent successes, the implemented solution requires three further improvements.

First, recombination was still detected in 3% of loci in the custom target dataset, even after alignment quality-control measures were performed. Additionally, the PHI test may be too conservative for alignments with low sequence diversity (Bruen et al., 2006). Therefore, given the low sequence diversity in the datasets used for this study, there is a risk of underestimating recombination using the PHI test. Over 27% of loci in the custom target dataset had too little sequence divergence for the PHI test to be used, as well as 25% of loci in the original dataset. The low sequence divergence observed in this study has also been observed in other studies of *Cirsium.* A phylogenetic study of North American *Cirsium* using Sanger sequencing of ITS and ETS regions of 18S-26S rDNA also recovered sequence data with low sequence divergence across taxa for such an ecologically diverse genus (Kelch & Baldwin, 2003). A potentially simple solution to the lack of sequence divergence observed in my datasets would be to extract intron-supercontig sequences from HybPiper for analysis. Inclusion of flanking regions in target-enrichment datasets has been helpful at solving relationships at shallow phylogenetic scales, in which low sequence divergence is expected (e.g., McKain et al., 2018; Herrando-Moraira et al., 2019). The inclusion of introns has additional utility in this dataset, having demonstrated that the over-alignment errors observed in the original dataset did not occur when introns are present.

Second, 38% of new targets triggered a default HybPiper paralog warning and were therefore removed from the analysis during alignment quality-control steps. Exclusion of loci flagged as paralogs is preferable to that locus containing undetected paralogs being incorporated into the dataset. However, the rate of paralog warnings triggered among loci assembled from new targets indicates that improvements can be made to the decision rules for creating the custom target file. One method to improve the custom target file would be to perform two or more iterations, in which loci still triggering paralog warnings after one iteration are further broken up into additional discrete loci representing each paralog. This process could be repeated until HybPiper paralog warnings no longer occur but would be time-intensive. However, this iterative process could be automated with custom scripts using the decision rules presented here.

Third, there may be additional as-yet-uninvestigated assembly errors that occurred. I examined some, but not all steps of the assembly process. There are multiple steps in the assembly process, and errors that may have occurred during earlier steps were not investigated. For example, many *de novo* gene assembly programs (including HybPiper) create a consensus sequence from raw sequence reads even when allelic variation is captured by the raw sequence reads (Kates et al., 2018; Andermann et al., 2019). In some instances, the assembled locus may be a combination of both alleles but is incorrectly interpreted as homozygous (Kates et al., 2018). Such assemblies represent artificial recombination occurring through a different mechanism than what was observed in this study and may confound biological conclusions of phylogenetic analyses when assemblies do not represent any in vivo allele. Attempts to assemble individual-allele sequences from target enrichment data have had mixed results, having phylogenetic utility in one study (Andermann et al., 2019), while having only a minor impact in another study (Kates et al., 2018).

Paralogs

The large proportion of identified undetected paralogs in the original *Cirsium* dataset indicate that multiple copies exist for many targeted genes, which is often, in plants, a result of polyploidy (i.e., homoeologous genes). However, there is not any evidence of recent genome duplication events in *Cirsium* or the Cardueae tribe (Huang et al., 2016). Cardueae was observed to have a low number of putative paralogs flagged by HybPiper in comparison to several other Compositae tribes (Herrando-Moraira et al., 2019). These observations beg the question of how often undetected paralogy occurs in other Compositae datasets generated using the Compositae1061 target enrichment kit. Furthermore, it would be interesting to examine how prevalent flagged and un-flagged paralogs are in datasets using the Angiosperms353 target enrichment kit (Johnson et al., 2019). Presumably, the more universal bait kit will have less specificity than the more taxon-specific probes and will therefore have even greater chance of recovering paralogs and undetected paralogs. However, Siniscalchi et al. (2021), in a direct comparison of data generated by Composiate1061 and Angiosperms353 kits, found paralogs to be more abundant in the Compositae1061 probe set. The deleterious effects of undiagnosed paralogs on phylogenetics datasets have been demonstrated (Koonin, 2005), but their inclusion, when properly diagnosed and parsed into orthologs, has also been shown to improve tree topologies in phylogenetic inference (Gardner et al., 2020; Frost & Lagomarsino, 2021). The effects of recombinant artifacts created during assembly processes, such as those observed in this dataset, have not been well studied in the context of phylogenomics.

HybPiper

Numerous assemblies were created using the original target file that likely do not represent any true *in vivo* allele. Given that HybPiper is a commonly used pipeline for target sequence capture in plant phylogenomics, I was surprised to find so many instances of gene assembly error. Perhaps the high frequency with which gene assembly errors occurred was unique to this *Cirsium* dataset, but I expect that these same assembly errors also occur in other datasets to some degree. While the tested solution reduced opportunities for artificial recombination and inconsistent contig selection to happen (i.e., separating competing contigs into discrete loci), there are other potential solutions that could be implemented. It was clear that HybPiper's default paralog warning threshold was not strict enough to correctly identify all instances of paralogy. Reducing the default length requirement for triggering a paralog warning would reduce the likelihood that paralogs go undetected in the dataset. Additionally, Exonerate only considers for extraction those contigs that have at least a 60% identity to the reference sequence. Increasing the default percent identity threshold would reduce the likelihood that multiple exon-contigs are recovered for a locus, thereby reducing the likelihood that paralogs go undetected in the dataset. Implementing changes to both of these settings would restrict the opportunity for both inconsistent contig selection as well as artificial recombination to occur.

The mechanism allowing for artificial recombination to occur between exon-contigs in cases of undetected paralogy merits further investigation. HybPiper's documentation does not explain how HybPiper performs when multiple exon contigs are recovered for a specimen that overlap by more than 20 bp but span less than 85% of length of the reference, and in which the longest contig does not completely subsume the range of all other contigs recovered for that specimen. My investigation shows that for these instances, HybPiper may or may not join contigs together to create an exon-supercontig. The assembly errors and inconsistencies discussed here should help direct further research into what improvements could be made to the algorithms and user-controlled settings within HybPiper.

Taxonomic Conclusions

The phylogenetic trees and network built from the final cleansed dataset were insufficient for conclusively resolving relationships among the sampled specimens. *Cirsium mohavense* (sensu lato) was polyphyletic in ML, parsimony, and coalescent inferences because outgroups used in the study did not form a bipartition in these same tree estimations. Furthermore, the network showed substantial reticulation among the sampled specimens, indicating that tree-based estimations may be inappropriate for the focal taxa.

The difficulty in resolving taxonomic relationships among North American *Cirsium* species was also reported by Ackerfield et al. (2020), in which many accepted *Cirsium* species, including *C. mohavense*, were resolved as polyphyletic. Therefore, this study provides further evidence that the currently accepted taxonomy of North American *Cirsium* needs substantial revision. There are several hypotheses as to the underlying causes of the difficulty in establishing correct species delimitations within *Cirsium*, such as a poor understanding of the morphological variation within and between species, phenotypic convergence, hybridization, and incipient speciation (Ackerfield et al., 2020). RAD-seq methods are a commonly used second-generation sequencing technique for obtaining population-level data, and may be more suitable for solving the relationships within *C. mohavense*, as well for all North American *Cirsium*, given the observed low sequence divergence in both this study and by Kelch and Baldwin (2003).

CONCLUSION

The results of this case study show the importance of viewing sequence data during quality-control and analysis. The automated alignment quality-control measures used in this dataset (i.e., MAFFT, trimAl, and TAPER) were insufficient at identifying, removing, and/or masking gene assembly errors and problematic regions resulting from over-alignment in the original dataset. But these errors were quite obvious when viewing the alignments. In the era of high-throughput sequencing in which hundreds to thousands of genes are sequenced, it is uncommon that researchers view their sequence data during phylogenetic studies. Viewing and critiquing aligned sequence data, which was demonstrated to be highly informative for this study, should become a standard practice in any phylogenetics or population genetics study.

As newer sequencing technology produces longer sequencing reads (i.e., third-generation sequencing), the difficulty of *de novo* gene assembly will decrease (Schatz et al., 2012). However, even once long-read sequencing is accurate enough to replace short-read sequencing approaches, the need to sample herbarium specimens will still exist. Making gene assemblies from highly fragmented sequence reads derived from herbarium specimens will remain a challenge, especially when these reads come from large genomes with many homoeologs and paralogs. Therefore, my observations and results presented here will continue to be relevant for gene assembly processes.

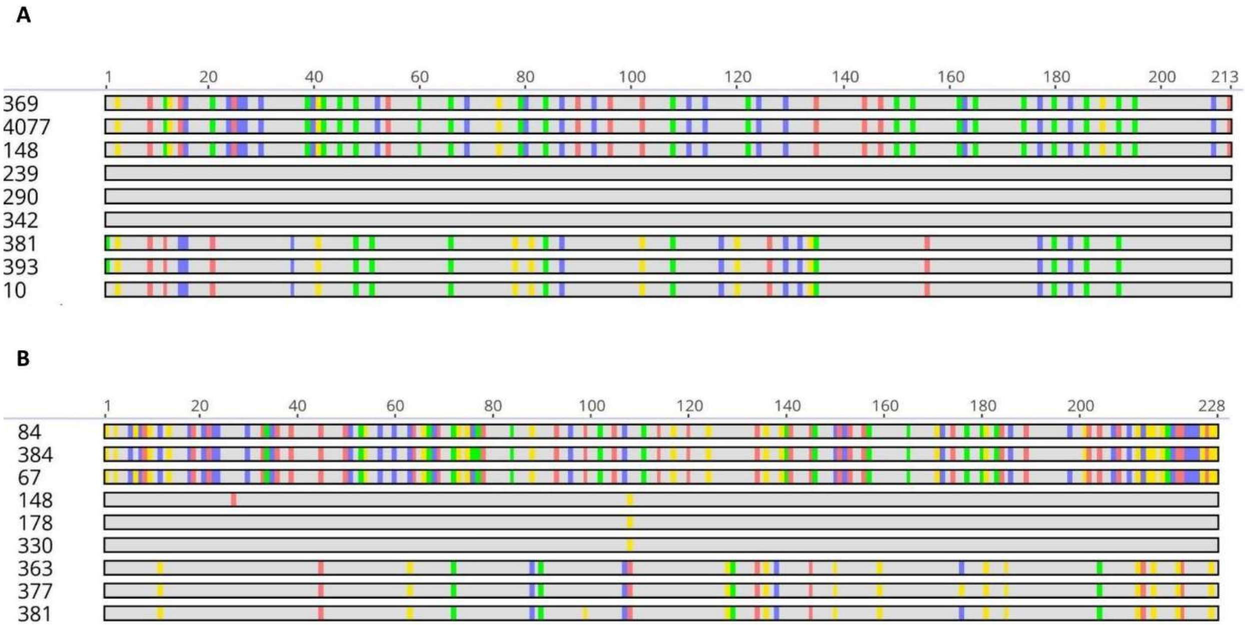

Figure 1. Panel A shows the final gene alignment (after alignment quality-control steps) for a gene that was not flagged as a paralog (At1g01050). Panel B shows the final gene alignment (after alignment quality-control steps) for a gene that triggered HybPiper's default paralog warning (At1g03220). For both gene alignments, an exemplar subset of the specimens is shown. The two alignments share a distinct pattern indicative of paralogy.

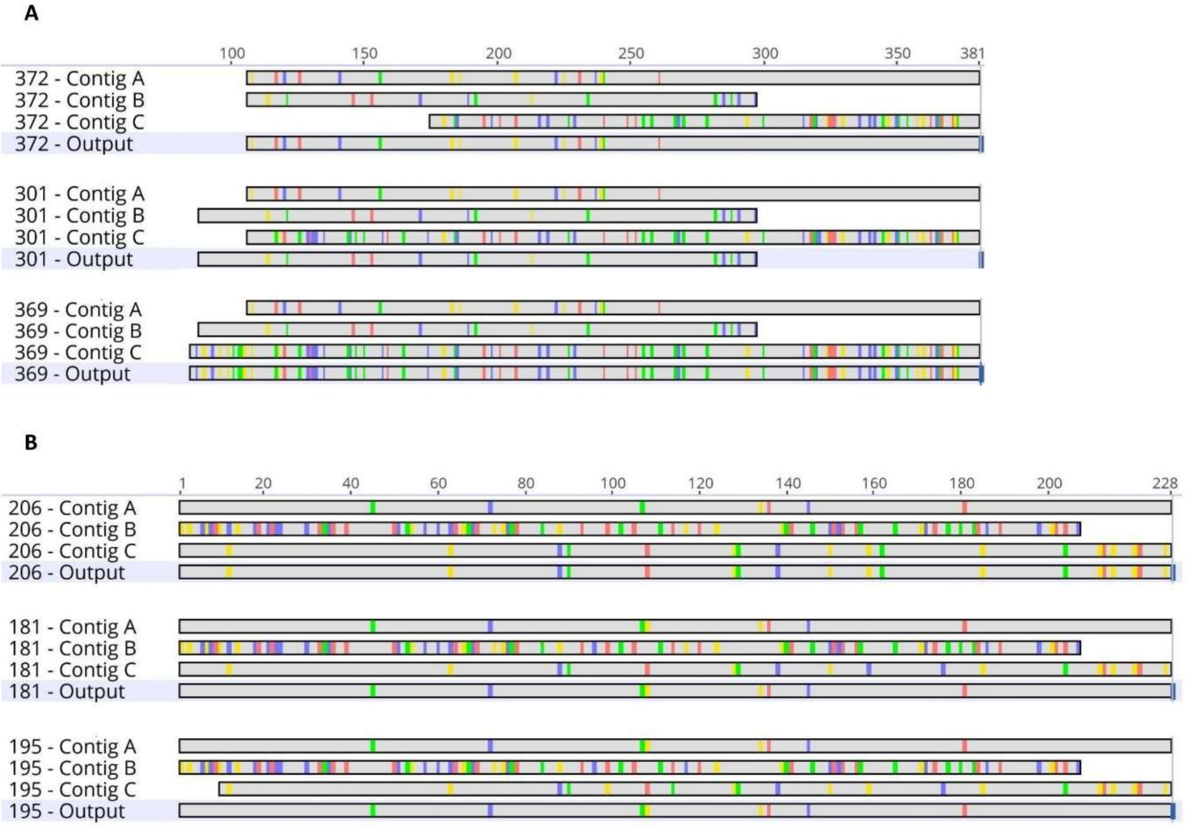

Figure 2. Panel A shows the exon contigs and final HybPiper output for a gene not flagged as including paralogs (At1g01050). Panel B shows the exon contigs and final HybPiper output for a gene that was flagged as including paralogs (At1g03220). Both panels show an exemplar subset of the specimens for simplicity. The pattern of competing contigs recovered for the unflagged gene was very similar to the flagged gene. Panel A shows inconsistent contig selection between specimens: the output sequence for specimen 372 was contig A, the output sequence for specimen 301 was contig B, and the output sequence for specimen 369 was output C. Inconsistent contig selection also occurred for flagged genes (panel B): the output sequence for specimens 181 and 195 is contig C and is contig A for specimen 206. This inconsistent contig selection was expected for flagged genes, which were excluded from the dataset before any analyses for this reason, but was an unexpected observation for unflagged genes.

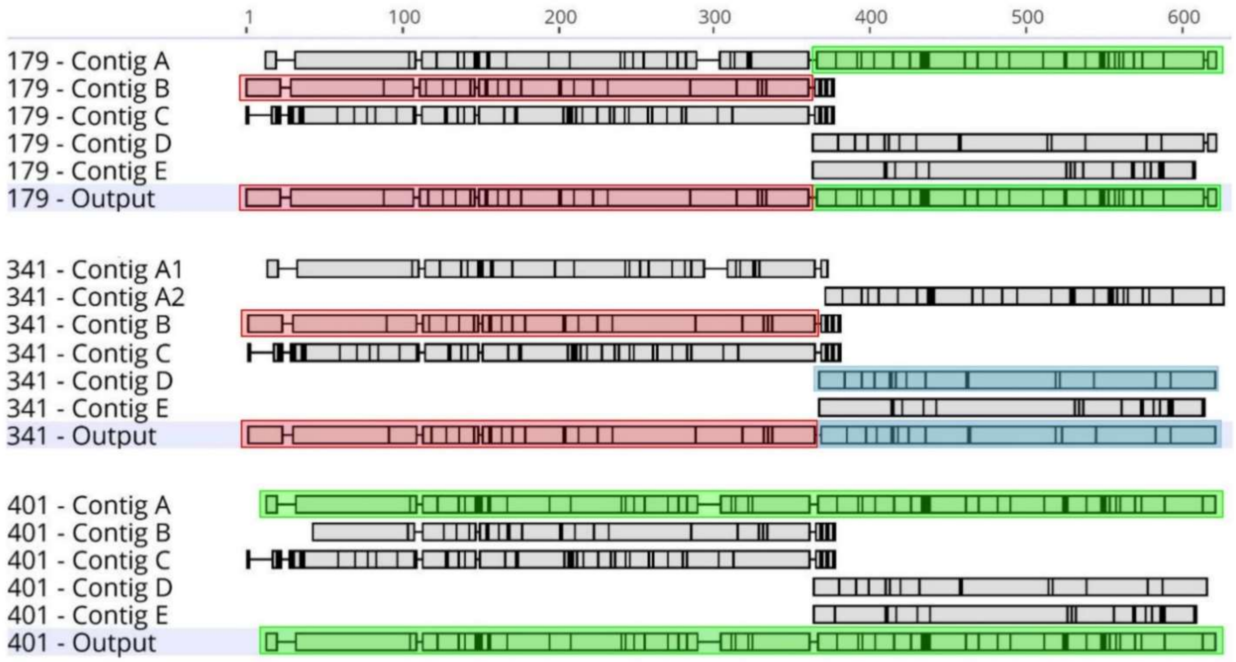

Figure 3. Artificial recombination in a gene with undetected paralogy (At1g32100). Alignment regions in black signify sequence bases that vary from the majority of specimens (i.e., the consensus sequence). In the upper specimen (179), the output sequence consisted of the 5’ end of contig B and the 3’ end of contig A. For the second specimen (341), the 5’ end of the output sequence came from contig B, but the 3’ end came from contig D. The two output sequences were not joined together in a consistent manner between specimens 179 and 341. However, the breakpoint for artificial recombination was consistent across these specimens, in contrast to the gene shown in figure 4. For the third specimen (401), HybPiper did not join contigs together, and the entirety of contig A was chosen for the output sequence.

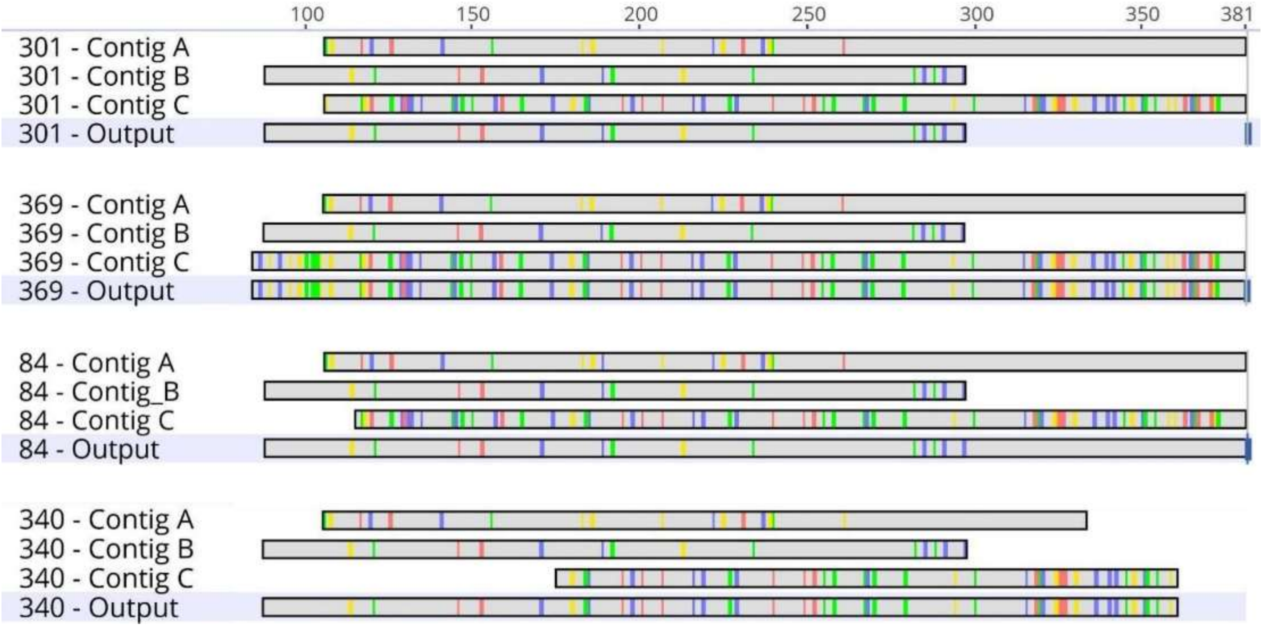

Figure 4. Co-occurrence of inconsistent contig selection and artificial recombination shown for gene At1g01050, with an exemplar subset of specimens. Specimens 301 and 369 showed inconsistent contig selection by HybPiper, while specimens 84 and 340 showed artificial recombination. The output sequence for specimen 84 consisted of contig B joined with the 5' end of contig A. The output sequence for specimen 340 consisted of contig C joined with the 5' end of contig B. The breakpoint for recombination was different between specimens 84 and 340, in contrast to the consistent recombination breakpoint observed in figure 3. In the case shown here, artificial recombination was observed among extensively overlapping contigs.

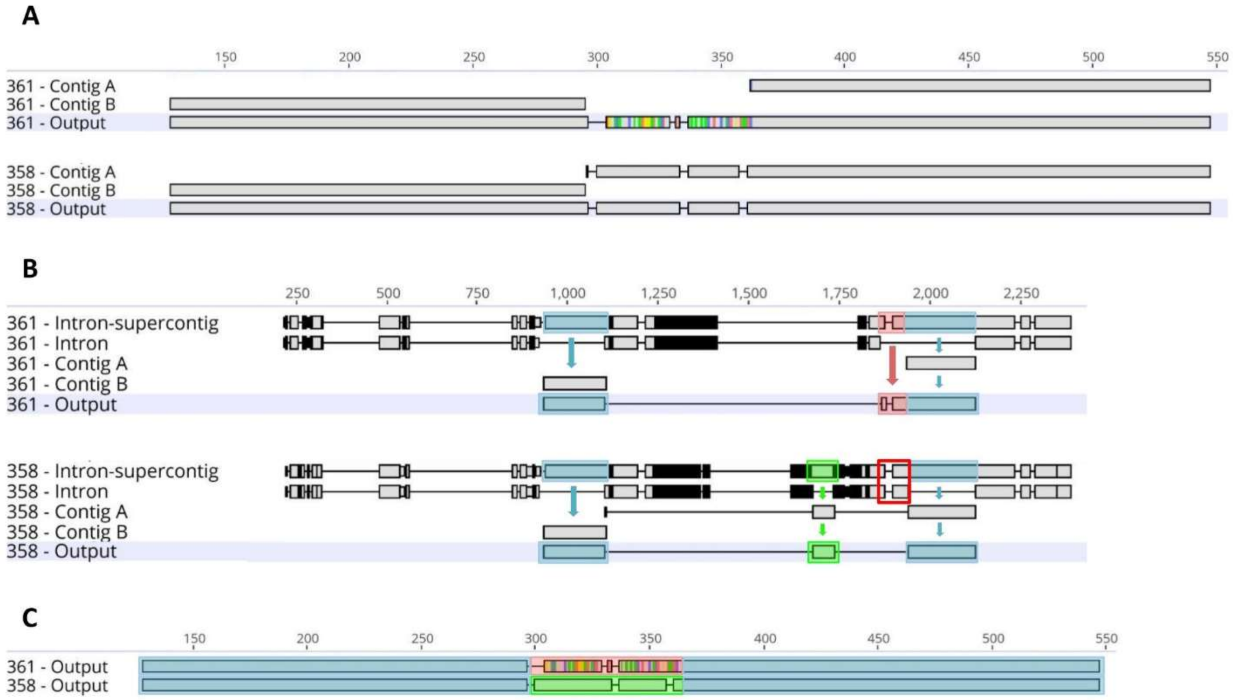

Figure 5. Panel A shows final exon-supercontig sequence data (gene At1g05720) in specimen 361 that does not originate from any recovered exon contig and is hypervariable when aligned with other specimens (only specimen 358 is shown for simplicity). Panel B shows an alignment with the intron-supercontig, intron, exon-contigs, and final output (exon-supercontig in this case) for the same two specimens. In panels A and B, alignment regions in black signify sequence bases that vary from the majority of specimens (i.e., the consensus sequence). The final output sequence for both specimens was simply the intron-supercontig minus the intron sequence. Regions in blue were consistently retained by Exonerate in both specimens (panel B) and were correctly aligned in the final gene alignment (panel C). The hypervariable region (highlighted in red) was determined by HybPiper to be exon sequence for specimen 361 but determined by HybPiper to be intron sequence for specimen 358, despite recovery of identical sequence in that region. Hence, this is a case of inconsistent intron determination. The region in green shows an exonic sequence region that was recovered for specimen 358, but not specimen 361. Panel C shows that the green and red alignment regions become mis-aligned by MAFFT without introns for reference, resulting in a hypervariable alignment region.

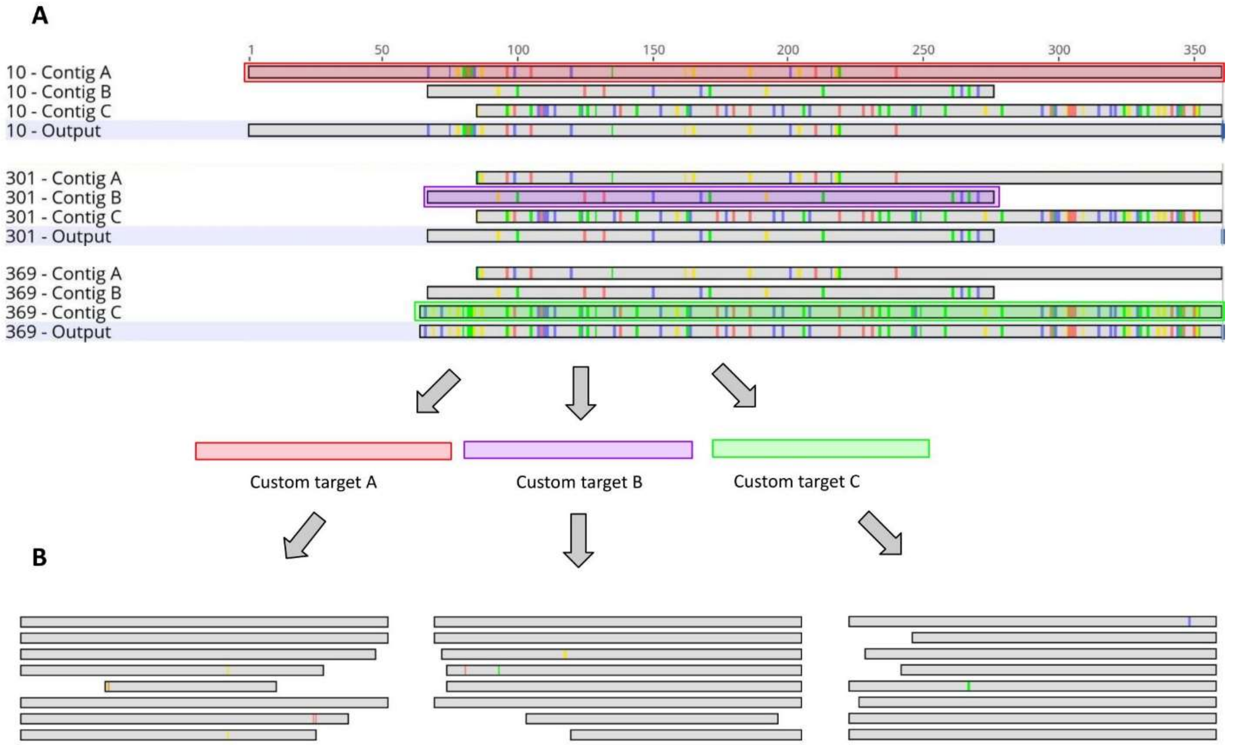

Figure 6. Diagram showing an example implementation of the proposed solution applied to gene At1g01050, and the outcome. The longest (or tied for longest) representative contig for each competing contig recovered across specimens was used as a new target reference sequence (Panel A). Assemblies using the custom targets were expected to separate the competing contigs into their own distinct gene assemblies. The proposed solution was successful at separating competing contigs for this test gene, based on multiple-sequence-alignments of the assembled new target loci (Panel B). An exemplar subset of the specimens is shown for each alignment.

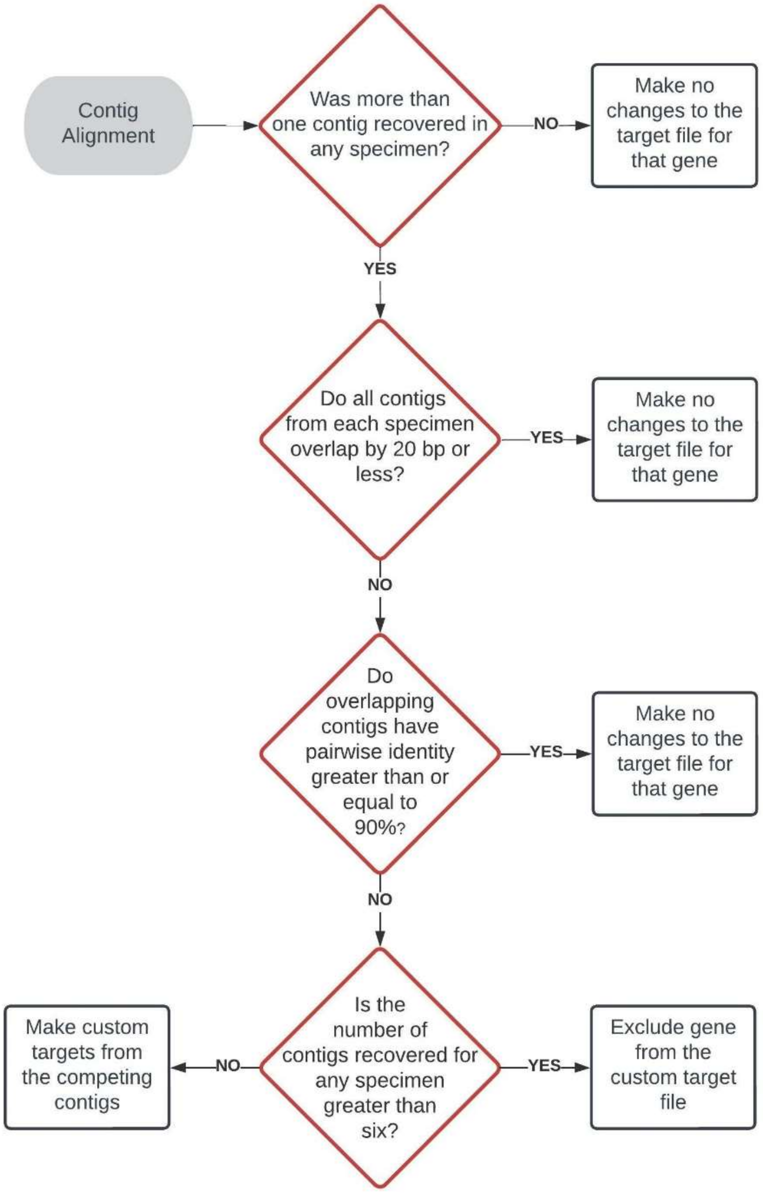
Contig Alignment
Was more than one contig recovered in any specimen?
NO
Make no changes to the target file for that gene
YES
Do all contigs from each specimen overlap by 20 bp or less?
YES
Make no changes to the target file for that gene
NO
Do overlapping contigs have pairwise identity greater than or equal to 90%?
YES
Make no changes to the target file for that gene
NO
Is the number of contigs recovered for any specimen greater than six?
NO
Make custom targets from the competing contigs
YES
Exclude gene from the custom target file

Figure 7. Flowchart of decision rules guiding the creation of the custom target file.

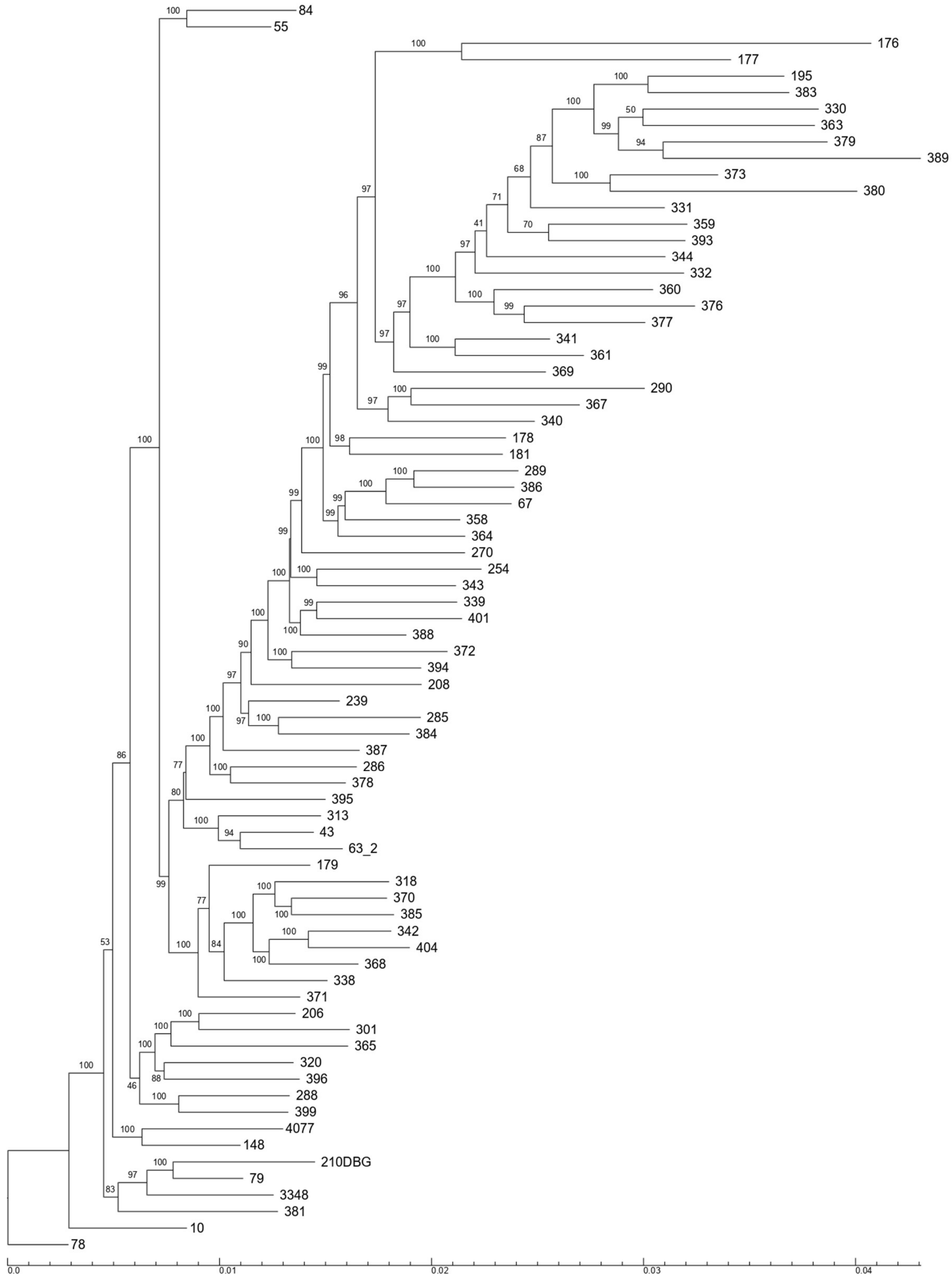

Figure 8. Maximum likelihood phylogeny inferred from the concatenated dataset produced using the original target file with bootstrap support shown above or below branches.

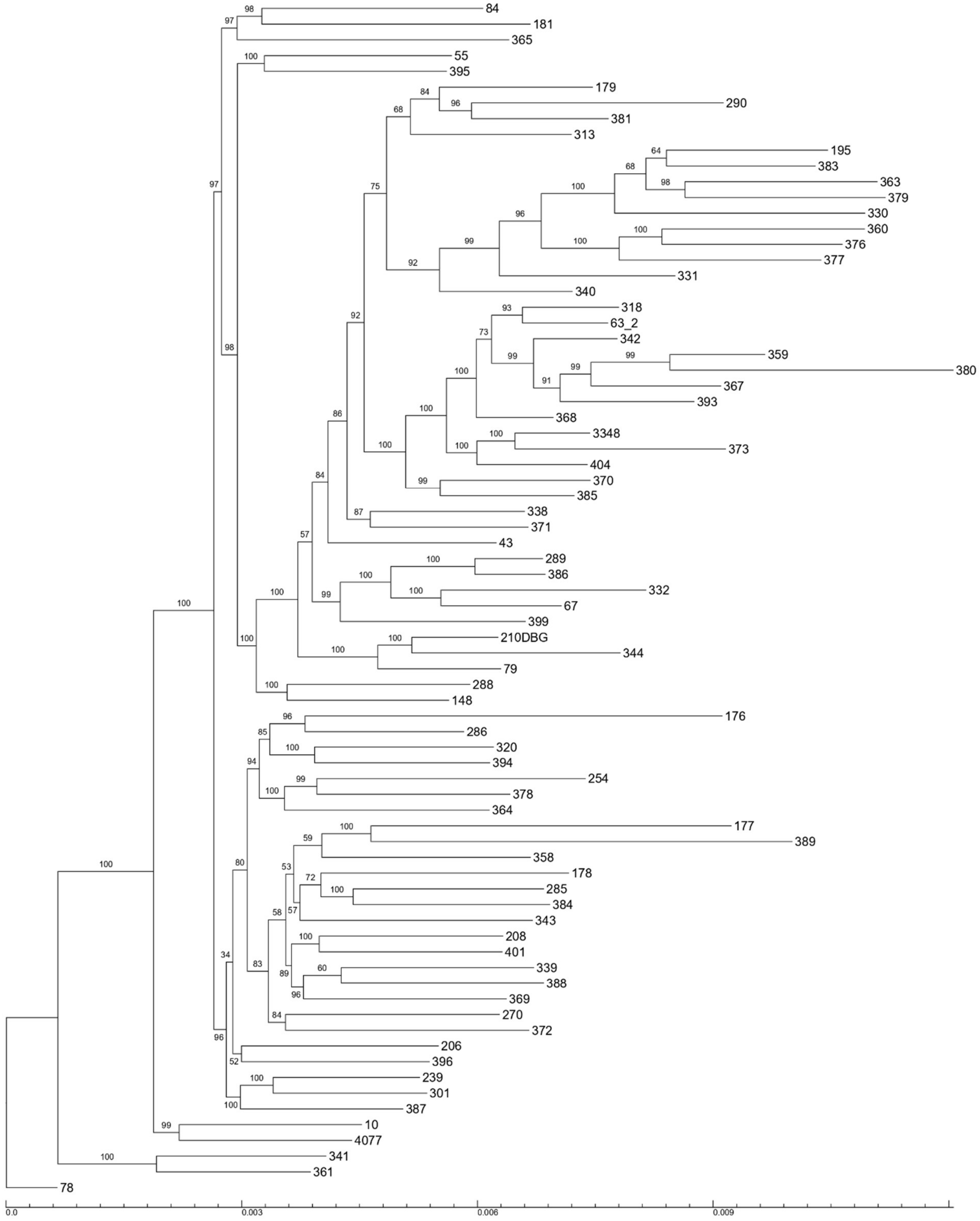

Figure 9. Maximum likelihood phylogeny inferred from the concatenated dataset producing using the custom target dataset with bootstrap support shown above or below branches.

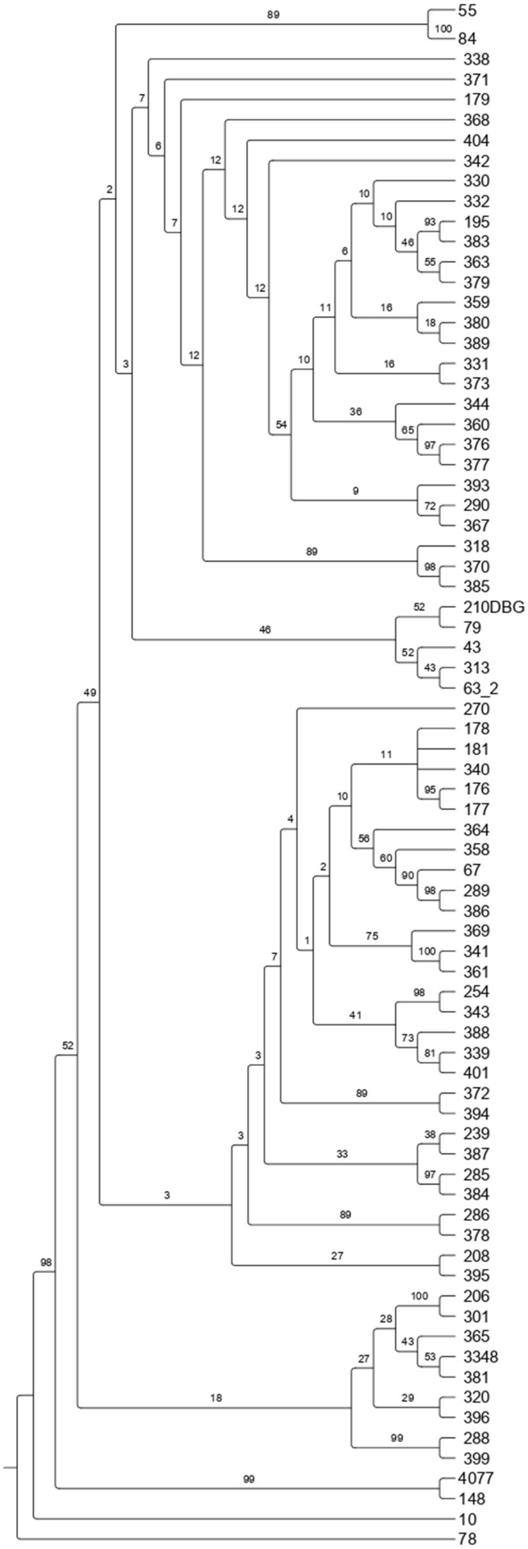

Figure 10. Parsimony phylogeny inferred from the concatenated dataset produced using the original target file with bootstrap support shown above or below branches.

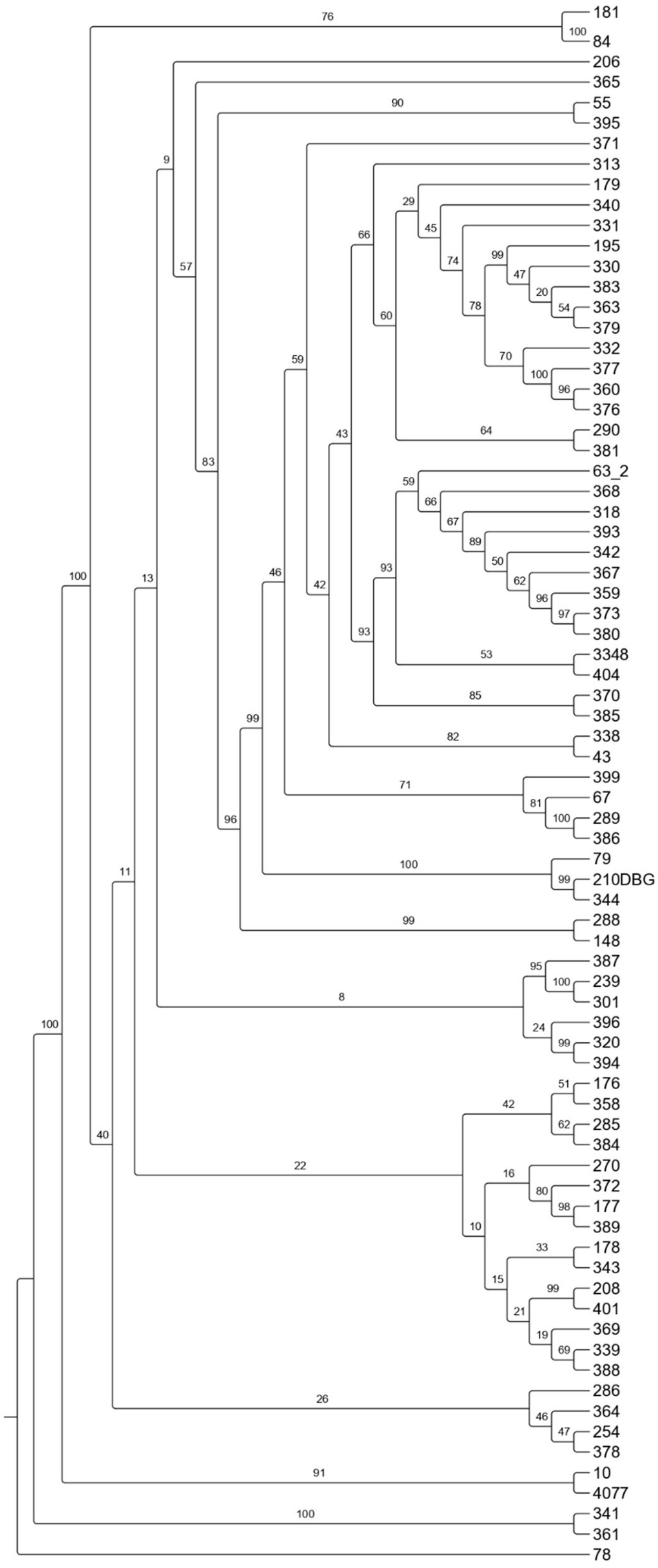

Figure 11. Parsimony phylogeny inferred from the concatenated dataset produced using the original target file with bootstrap support shown above or below branches.

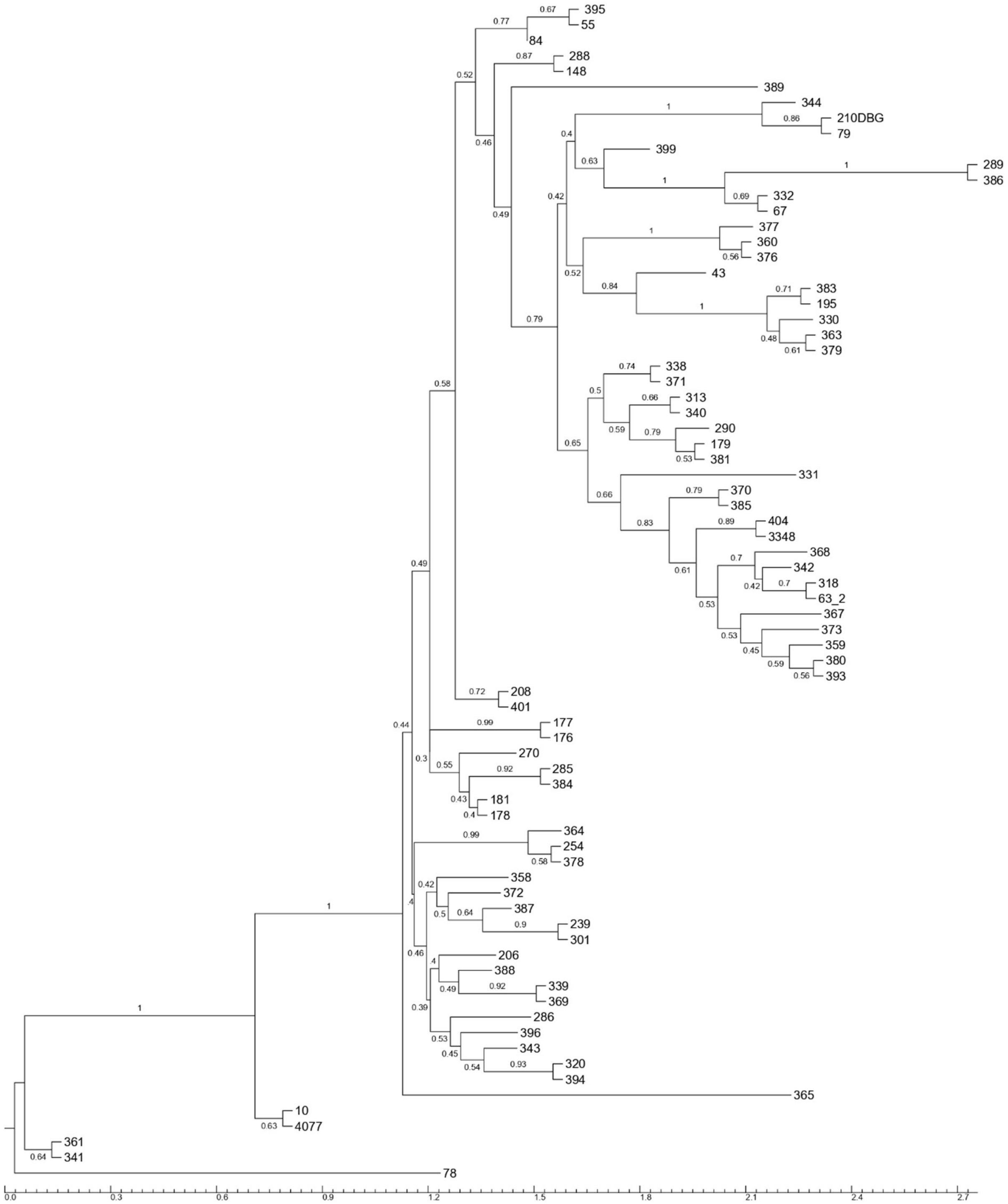

Figure 12. Non-parametric ASTRAL species tree inferred from parsimony strict-consensus gene trees produced using the original target file with local posterior probability (LPP) shown above or below branches.

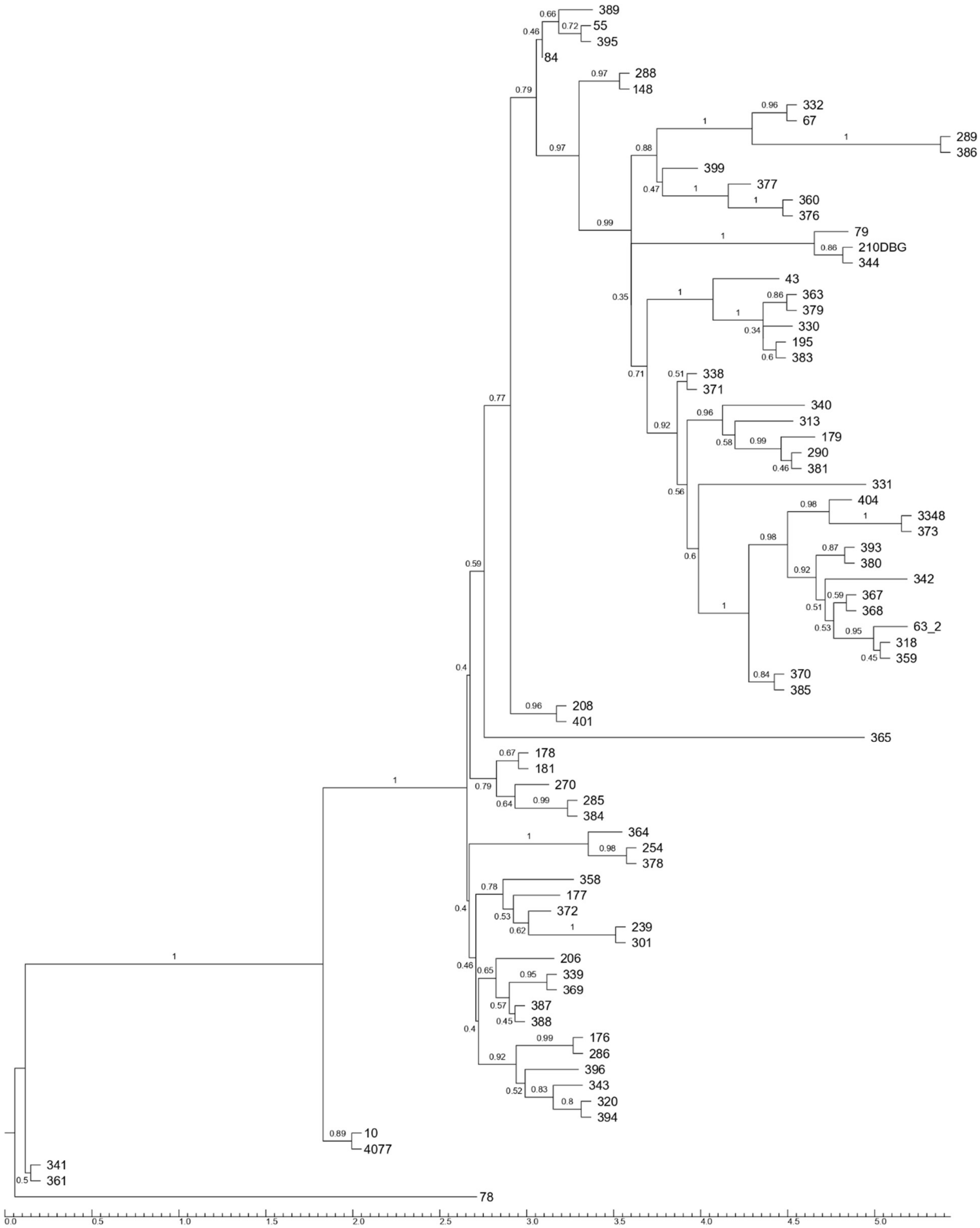

Figure 13. Non-parametric ASTRAL species tree inferred from parsimony gene trees produced using the custom target file with local posterior probability (LPP) shown above or below branches.

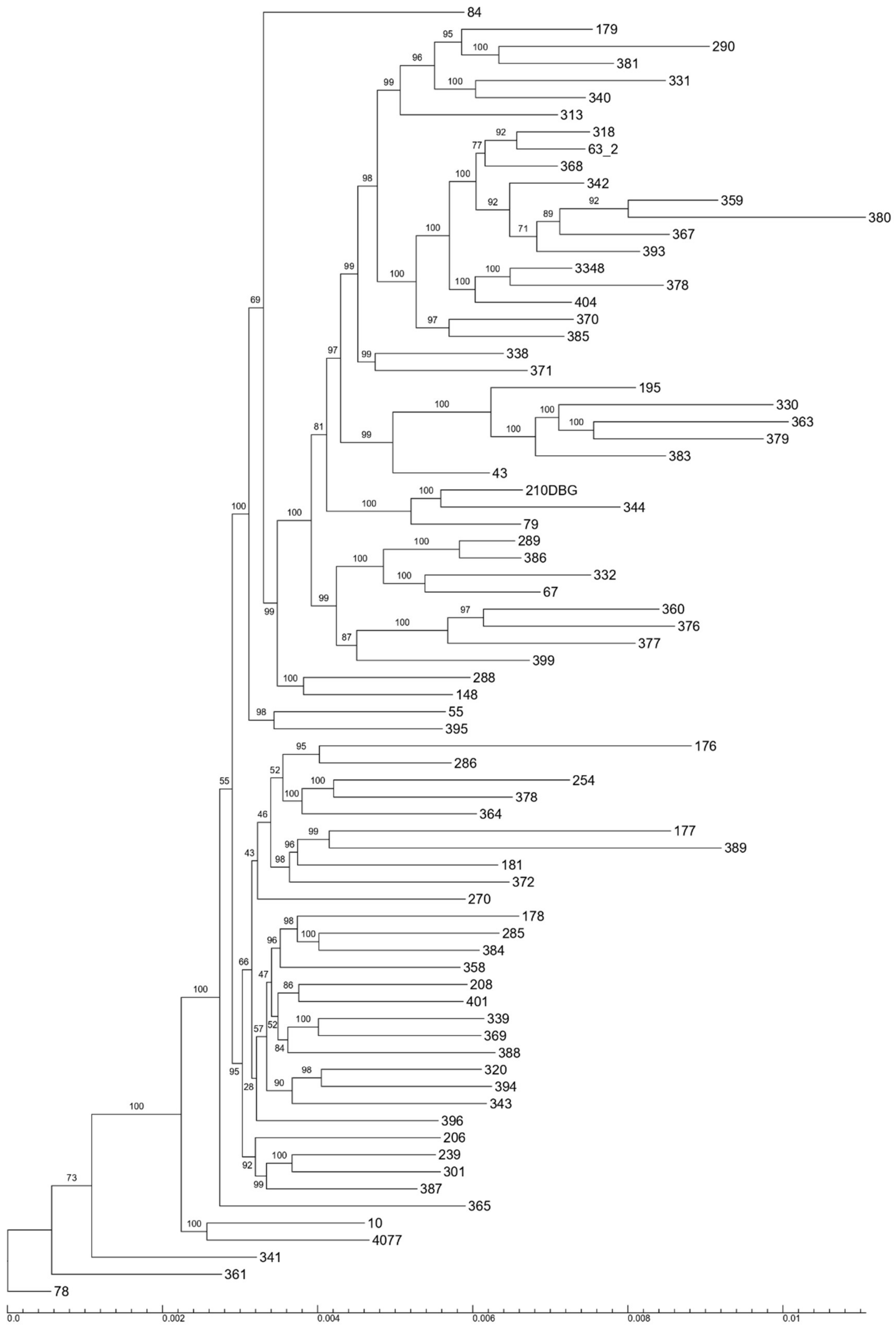

Figure 14. Maximum likelihood phylogeny inferred from the concatenated dataset producing using the final cleansed dataset with bootstrap support shown above or below branches.

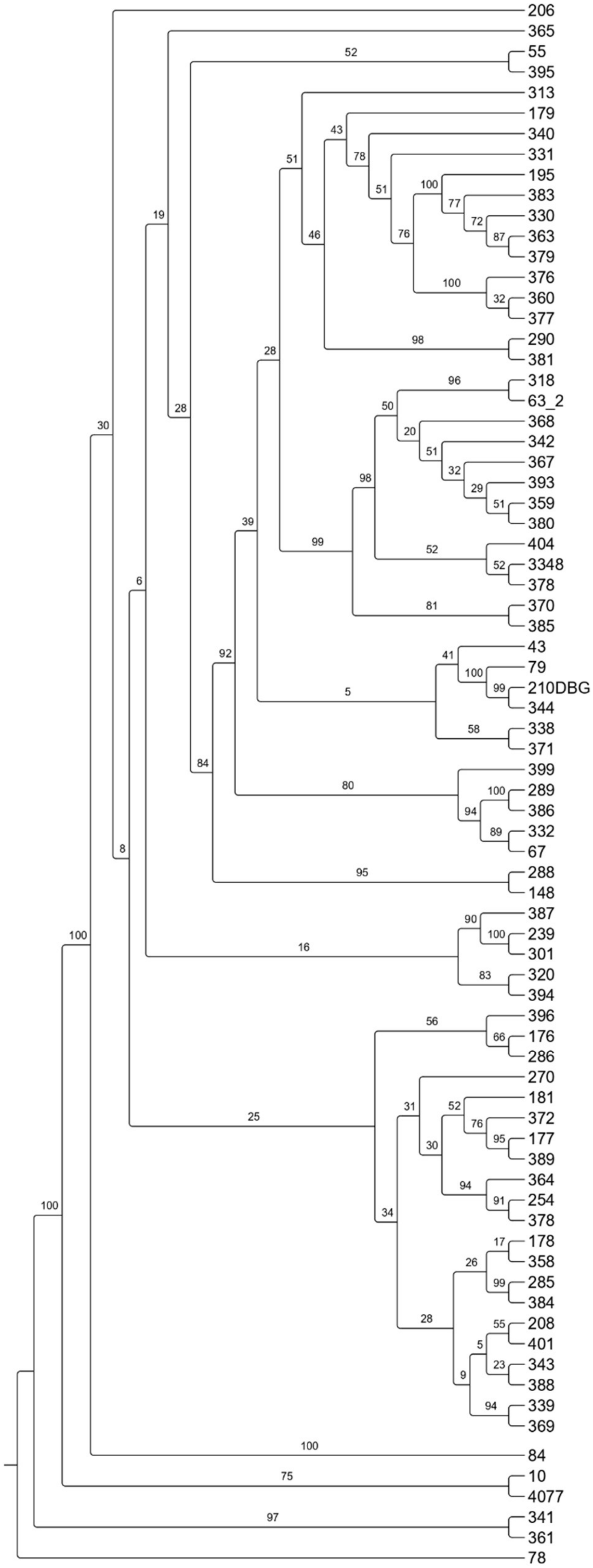

Figure 15. Parsimony phylogeny inferred from the concatenated final cleansed dataset with bootstrap support shown above or below branches.

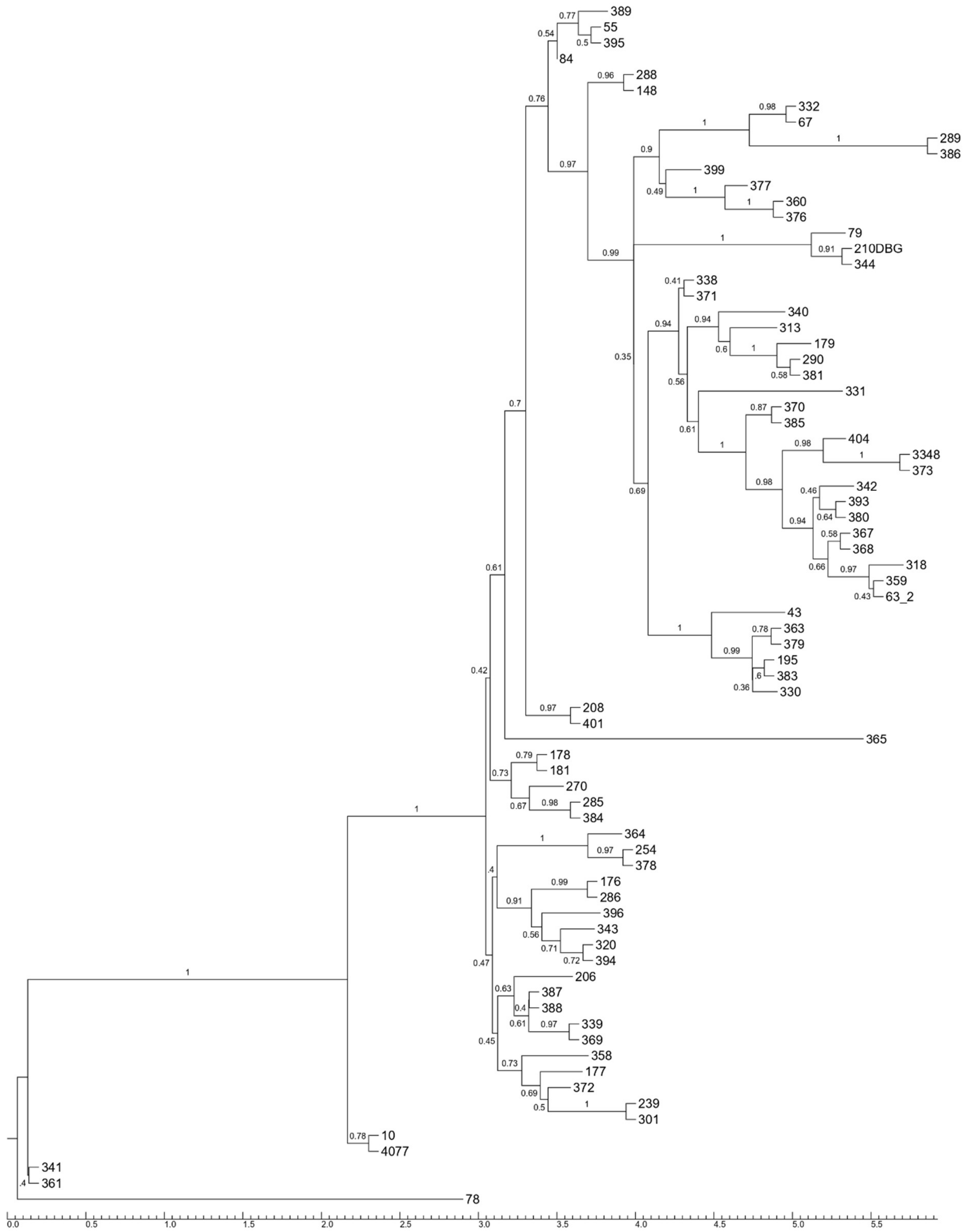

Figure 16. Non-parametric ASTRAL species tree inferred from parsimony gene trees in the final cleansed dataset with local posterior probability (LPP) shown above or below branches.

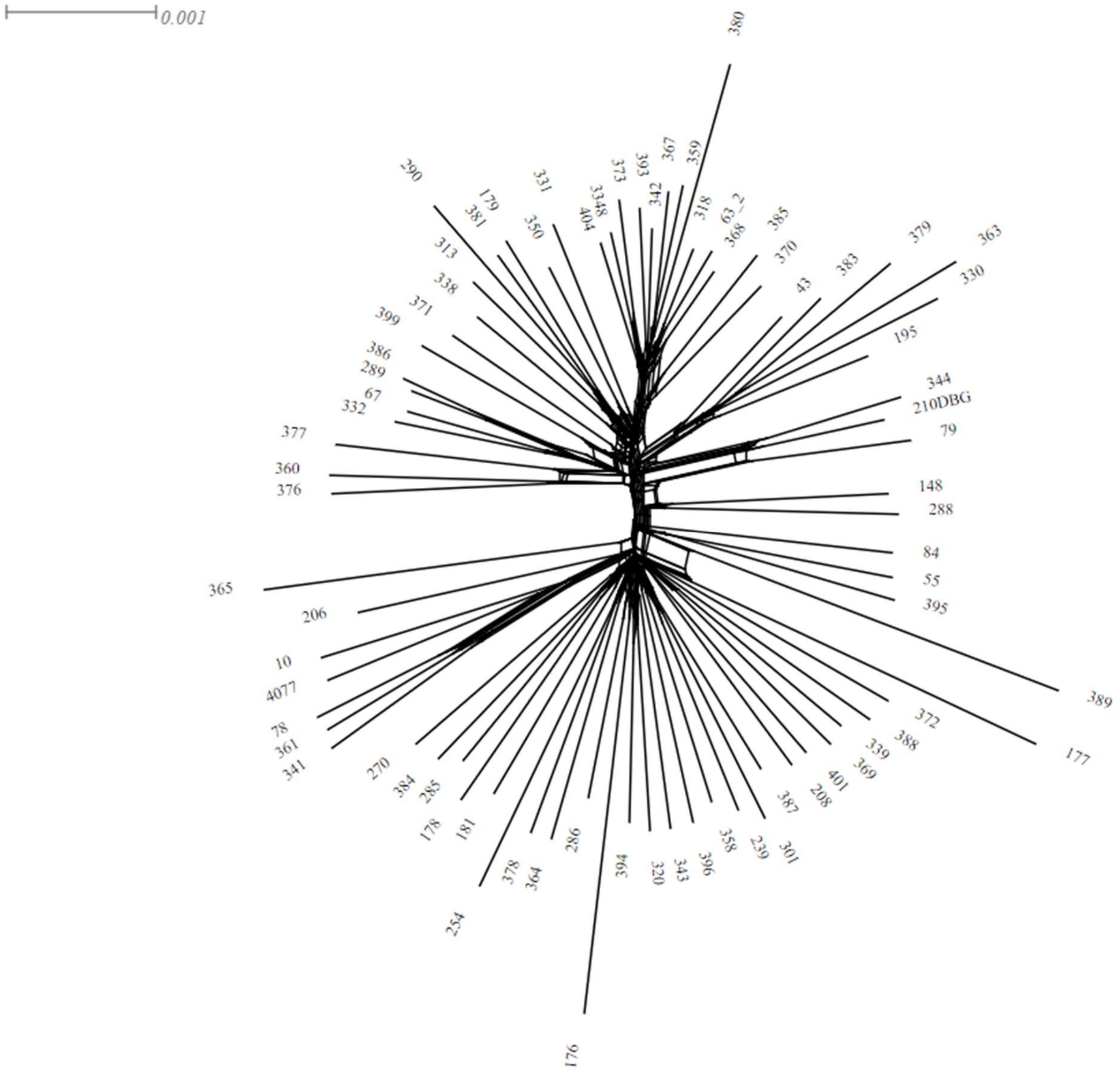

Figure 17. The NeighborNet network for the final cleansed dataset does not show obvious clustering of any groups of specimens that appeared highly distinct from any other groups of specimens. Rather, the network showed a general pattern of short boxy (i.e., conflicting characters in a hierarchical context) internal branches in comparison to long terminal branches, similar to the trees inferred by ML methods. For example, there was obvious conflicting hierarchical signal for specimens 148, 288, and 389. While the network shows evidence of a reticulate relationship among specimens in the study, the scale is dominated by the long terminal branches resulting from inferred autapomorphies.

Table 1. Results of target enrichment extraction from using the original and custom target files.

Target File Type	Loci in Target File	Recovered Loci	Mean Target Length (bp)	Mean Recovered Contig Length (bp)	Mean Contigs per Locus	Mean Contigs per Paralog	Total Paralog	Mean Paralogs per Specimen
Original	1061	1030	404	307	2.5	3.4	272	121
Custom	1593	1564	330	275	1.4	1.4	382	41

Table 2. Summary statistics of the PHI test across gene assemblies generated using the original and custom target files, after adjusting p-values for multiple tests using the FDR method.

Target File Type	Total Loci with No Result	Total Recombinant Loci	Proportion Recombinant Loci	Total Loci Passed or No Result	Proportion of Loci Passed or No Result
Original	228	132	0.18	612	0.82
Custom	432	38	0.03	1124	0.97

Table 3. Concatenated matrix summary statistics, and comparison of phylogenetic inferences by target file.

Dataset	Total Loci	Total Characters	Parsimony Informative Characters	ML Branch Length Sum	ML Mean Bootstrap	Parsimony Mean Bootstrap	RI
Original	744	265,639	14,604	0.58	81.5	51.4	0.36
Custom	1162	371,910	8,482	0.21	89.2	66.0	0.46
Final Cleansed	1124	354,609	7,417	0.19	90.0	61.3	0.43

Table 4. Comparison of coalescent results by target file.

Dataset	Total Loci	NQS	Mean LPP
Original	744	0.52	0.65
Custom	1162	0.66	0.76
Final Cleansed	1124	0.67	0.76

REFERENCES

Ackerfield, J. R., Keil, D. J., Hodgson, W. C., Simmons, M. P., Fehlberg, S. D., & Funk, V. A. (2020). Thistle be a mess: Untangling the taxonomy of *Cirsium* (Cardueae: Compositae) in North America. *Journal of Systematics and Evolutio*n, 58(6), 881–912. https://doi.org/https://doi.org/10.1111/jse.12692

Alexander, P., Govindarajulu, R., Bacon, C., & Bailey, C. D. (2007). Recovery of plant DNA using a reciprocating saw and silica-based columns. *Molecular Ecology Notes*, 7, 5–9. https://doi.org/10.1111/j.1471-8286.2006.01549.x

Andermann, T., Fernandes, A. M., Olsson, U., Töpel, M., Pfeil, B., Oxelman, B., Aleixo, A., Faircloth, B. C., & Antonelli, A. (2019). Allele phasing greatly improves the phylogenetic utility of ultraconserved elements. *Systematic Biology*, 68(1), 32–46. https://doi.org/10.1093/sysbio/syy039

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V, Sirotkin, A. V, Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, 19(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Borowiec, M. L. (2016). AMAS: A fast tool for alignment manipulation and computing of summary statistics. *PeerJ*, 4, e1660. https://doi.org/10.7717/peerj.1660

Bruen, T. (2005). PhiPack: PHI test and other tests of recombination.

Bruen, T., Philippe, H., & Bryant, D. (2006). A simple and robust statistical test for detecting the presence of recombination. *Genetics*, 172, 2665–2681. https://doi.org/10.1534/genetics.105.048975

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15), 1972–1973. https://doi.org/10.1093/bioinformatics/btp348

de Queiroz, K. (1998). The general lineage concept of species, species criteria, and the process of speciation. In D. J. Howard & S. H. Berlocher (Eds.), *Endless Forms: Species and Speciation*. Oxford University Press.

Doyle, J. J. (1992). Gene trees and species trees: Molecular systematics as one-character taxonomy. *Systematic Botany*, 17(1), 144–163. https://doi.org/10.2307/2419070

Faircloth, B. C. (2016). PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics*, 32(5), 786–788. https://doi.org/10.1093/bioinformatics/btv646

Farris, J. S. (1989). The retention index and the rescaled consistency index. *Cladistics*, 5(4), 417–419. https://doi.org/https://doi.org/10.1111/j.1096-0031.1989.tb00573.x

Fér, T., & Schmickl, R. E. (2018). HybPhyloMaker: Target enrichment data analysis from raw reads to species trees. *Evolutionary Bioinformatics*, 14, 1176934317742613. https://doi.org/10.1177/1176934317742613

Frost, L., & Lagomarsino, L. (2021). More-curated data outperforms more data: Treatment of cryptic and known paralogs improves phylogenomic analysis and resolves a northern Andean origin of *Frezeria* (Pentaphylacaceae). BioRxiv, 2021.07.01.450750. https://doi.org/10.1101/2021.07.01.450750

Gardner, E. M., Johnson, M. G., Pereira, J. T., Puad, A. S. A., Arifiani, D., Wickett, N. J., & Zerega, N. J. C. (2020). Paralogs and off-target sequences improve phylogenetic resolution in a densely-sampled study of the breadfruit genus (*Artocarpus*, Moraceae). *Systematic Biology*, 70(3), 558–575. https://doi.org/10.1093/sysbio/syaa073

Goloboff, P. A., Farris, J. S., & Nixon, K. C. (2008). TNT, a free program for phylogenetic analysis. *Cladistics*, 24(5), 774–786. https://doi.org/https://doi.org/10.1111/j.1096-0031.2008.00217.x

Hale, H., Gardner, E. M., Viruel, J., Pokorny, L., & Johnson, M. G. (2020). Strategies for reducing per-sample costs in target capture sequencing for phylogenomics and population genomics in plants. *Applications in Plant Sciences*, 8(4), e11337. https://doi.org/10.1002/aps3.11337

Herrando-Moraira, S., Calleja, J. A., Carnicero, P., Fujikawa, K., Galbany-Casals, M., Garcia-Jacas, N., Im, H. T., Kim, S. C., Liu, J. Q., López-Alvarado, J., López-Pujol, J., Mandel, J. R., Massó, S., Mehregan, I., Montes-Moreno, N., Pyak, E., Roquet, C., Sáez, L., Sennikov, A., … Vilatersana, R. (2018). Exploring data processing strategies in NGS target enrichment to disentangle radiations in the tribe Cardueae (Compositae). *Molecular Phylogenetics and Evolution*, 128, 69–87. https://doi.org/10.1016/J.YMPEV.2018.07.012

Herrando-Moraira, S., Calleja, J. A., Galbany-Casals, M., Garcia-Jacas, N., Liu, J.-Q., López-Alvarado, J., López-Pujol, J., Mandel, J. R., Massó, S., Montes-Moreno, N., Roquet, C., Sáez, L., Sennikov, A., Susanna, A., & Vilatersana, R. (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. https://doi.org/https://doi.org/10.1016/j.ympev.2019.05.001

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: Improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution*, 35(2), 518–522. https://doi.org/10.1093/molbev/msx281

Hodgson, W. C., & Rink, G. (2018). A new species of *Cirsium* (Asteraceae) from northwestern Arizona. *Unpublished manuscript*.

Huang, C.-H., Zhang, C., Liu, M., Hu, Y., Gao, T., Qi, J., & Ma, H. (2016). Multiple polyploidization events across Asteraceae with two nested events in the early history revealed by nuclear phylogenomics. *Molecular Biology and Evolution,* 33(11), 2820–2835. https://doi.org/10.1093/molbev/msw157

Huson, D. H., & Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23(2), 254–267. https://doi.org/10.1093/molbev/msj030

Johnson, M. G., Gardner, E. M., Liu, Y., Medina, R., Goffinet, B., Shaw, A. J., Zerega, N. J. C., & Wickett, N. J. (2016). HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in Plant Sciences*, 4(7), apps.1600016. https://doi.org/10.3732/apps.1600016

Johnson, M. G., Pokorny, L., Dodsworth, S., Botigué, L. R., Cowan, R. S., Devault, A., Eiserhardt, W. L., Epitawalage, N., Forest, F., Kim, J. T., Leebens-Mack, J. H., Leitch, I. J., Maurin, O., Soltis, D. E., Soltis, P. S., Wong, G. K., Baker, W. J., & Wickett, N. J. (2019). A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. *Systematic Biology*, 68(4), 594–606. https://doi.org/10.1093/sysbio/syy086

Kates, H. R., Johnson, M. G., Gardner, E. M., Zerega, N. J. C., & Wickett, N. J. (2018). Allele phasing has minimal impact on phylogenetic reconstruction from targeted nuclear gene sequences in a case study of *Artocarpus*. *American Journal of Botany*, 105(3), 404–416. https://doi.org/https://doi.org/10.1002/ajb2.1068

Katoh, K., & Standley, D. M. (2013). MAFFT: Iterative refinement and additional methods. *Multiple Sequence Alignment Methods*, 1079, 131–146. https://doi.org/10.1007/978-1-62703-646-7_8

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., & Drummond, A. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12), 1647–1649. https://doi.org/10.1093/bioinformatics/bts199

Keil, D. J. (2006). *Cirsium*. In Flora of North America Editorial Committee (Eds.), *Flora of North America North of Mexico*, 95–164. Oxford University Press.

Kelch, D. G., & Baldwin, B. G. (2003). Phylogeny and ecological radiation of New World thistles (*Cirsium*, Cardueae - Compositae) based on ITS and ETS rDNA sequence data. *Molecular Ecology*, 12(1), 141–151. https://doi.org/10.1046/j.1365-294X.2003.01710.x

Landan, G., & Graur, D. (2007). Heads or tails: A simple reliability check for multiple sequence alignments. *Molecular Biology and Evolution*, 24(6), 1380–1383. https://doi.org/10.1093/molbev/msm060

Mandel, J. R., Dikow, R. B., Funk, V. A., Masalia, R. R., Staton, S. E., Kozik, A., Michelmore, R. W., Rieseberg, L. H., & Burke, J. M. (2014). A target enrichment method for gathering phylogenetic information from hundreds of loci: an example from the Compositae. *Applications in Plant Sciences*, 2(2), 1300085. https://doi.org/https://doi.org/10.3732/apps.1300085

McKain, M. R., Johnson, M. G., Uribe-Convers, S., Eaton, D., & Yang, Y. (2018). Practical considerations for plant phylogenomics. *Applications in Plant Sciences*, 6(3), e1038. https://doi.org/10.1002/aps3.1038

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, 37(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Nevada Natural Heritage Program. (2022). At-risk plant and animal tracking list. https://heritage.nv.gov/documents/ndnh-current-tracking-list

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32(1), 268–274. https://doi.org/10.1093/molbev/msu300

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.r-project.org/

Schatz, M. C., Delcher, A. L., & Salzberg, S. L. (2010). Assembly of large genomes using second-generation sequencing. *Genome Research*, 20(9), 1165–1173. https://doi.org/10.1101/gr.101360.109

Schatz, M. C., Witkowski, J., & McCombie, W. R. (2012). Current challenges in de novo plant genome sequencing and assembly. *Genome Biology*, 13(4), 243. https://doi.org/10.1186/gb4015

Simmons, M. P., & Gatesy, J. (2021). Collapsing dubiously resolved gene-tree branches in phylogenomic coalescent analyses. *Molecular Phylogenetics and Evolution*, 158, 107092. https://doi.org/https://doi.org/10.1016/j.ympev.2021.107092

Siniscalchi, C. M., Hidalgo, O., Palazzesi, L., Pellicer, J., Pokorny, L., Maurin, O., Leitch, I. J., Forest, F., Baker, W. J., & Mandel, J. R. (2021). Lineage-specific vs. universal: A comparison of the Compositae1061 and Angiosperms353 enrichment panels in the sunflower family. *Applications in Plant Sciences*, 9(7), n/a. https://doi.org/10.1002/aps3.11422

Thiers, B. (2016). Index herbariorum: A global directory of public herbaria and associated staff. New York Botanical Garden’s Virtual Herbarium. Available from http://sweetgum.nybg.org/ih

Utah Native Plant Society. (2022). Utah rare plant guide. https://www.utahrareplants.org/rpg_species.html

Welsh, S. L. (1982). New taxa of thistles (*Cirsium*; Asteraceae) in Utah. *Great Basin Naturalist*, 49(2), 199–202.

Welsh, S. L. (1983). Utah flora: Compositae (Asteraceae). *Great Basin Naturalist*, 43(2), 245–255.

Yuan, H., Atta, C., Tornabene, L., & Li, C. (2019). Assexon: Assembling exon using gene capture data. *Evolutionary Bioinformatics Online*, 15, 1176934319874792. https://doi.org/10.1177/1176934319874792

Zhang, C., Rabiee, M., Sayyari, E., & Mirarab, S. (2018). ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, 19(6), 153. https://doi.org/10.1186/s12859-018-2129-y

Zhang, C., Zhao, Y., Braun, E. L., & Mirarab, S. (2021). TAPER: Pinpointing errors in multiple sequence alignments despite varying rates of evolution. *Methods in Ecology and Evolution*, 12(11), 2145–2158. https://doi.org/https://doi.org/10.1111/2041-210X.13696

APPENDIX

Appendix A. Voucher information and herbarium index (Thiers, 2016) for all 76 specimens sampled in this study.
*Denotes samples sequenced in the first batch of library preparation and sequencing.

Bioinformatics Prefix	Voucher
10*	U.S.A.: Colorado. Delta Co.: Grand Mesa, Petrie Mesa ca 7 air mi NNW of Delta, 2 Jun 2011, L Brummer 3594 (RM)
43*	U.S.A.: California. San Bernardino Co.: Bighorn Mts Wilderness, unnamed spring WSW of Mount Spring, off Mound Spring Rd RC2248, 30 Jul 2014, S DeGroot 7219 (RSA)
55*	U.S.A.: Arizona. Gila Co.: Carol Site, off State Rt 77, 17 Jul 2003, D Damrel V 830 (ASU)
63_2*	U.S.A.: Arizona. Coconino Co.: Grand Canyon, Three Springs, 17 May 2005, W Hodgson 19727 (DES)
67*	U.S.A.: Nevada. Nye Co.: Big Smokey Valley 1.5 mi E of 376 on Northumberland Mine Rd., 18 Jul 2001, B Niell s.n. (RENO)
78*	U.S.A.: Nevada. Pershing Co.: Humboldt Range, NW end of Spring Valley, N of Spring Valley Pass, 25 Jun 2011, A Tiehm 16386 (RENO)
79*	U.S.A.: Utah. Washington Co.: northern saline seep W of Grapevine Pass Wash, ca. 1.5 mi NE of Washington, 18 May 2004, J Alexander 1690 (RENO)
84*	U.S.A.: Arizona. Coconino Co.: Coconino National Forest N rim of West Fork Canyon, tip of Harding Point, 14 Jun 2001, E Gilbert 712 (ASU)
148*	U.S.A.: Nevada. Clark Co.: Toiyabe National Forest, Spring Mountains, 17 Aug 2002, D Keil 29912C (OBI)
176	U.S.A.: Nevada. White Pine Co.: Spring Valley, 11 Jul 2016, A Tiehm 17557 (RENO)
177	U.S.A.: Nevada. Nye Co.: Cold Springs, 11 Jun 2017, M Baker 19076 (RENO)
178	U.S.A.: Nevada. Lincoln Co.: Pahranagat Valley, Crystal Springs, 05 Sep 2005, A Tiehm 15078 (RENO)
179	U.S.A.: Nevada. Clark Co.: Gold Butte National Monument, Little Finland, 24 Mar 2018, D Gentilcore 1453 (RENO)
181	U.S.A.: Nevada. Nye Co.: Railroad Valley, seepage, 06 Jul 2008, A Tiehm 15663 (RENO)
195	U.S.A.: Arizona. Mohave Co.: Santa Maria Springs, 10 Dec 2016, W Hodgson 31126 (DES)
206	U.S.A.: Arizona. La Paz Co: Grapevine Springs, 10 Dec 2016, W Hodgson 31128 (DES)
208	U.S.A.: Arizona. La Paz Co: Grapevine Springs, 10 Dec 2016, W Hodgson 31130 (DES)

210DBG*	U.S.A.: Utah. Washington Co.: Seep along road ca 0.5 mi east of Danish Ranch, 19 Aug 1986, R Warrick 2961 (BRY)
239	U.S.A.: Arizona. La Paz Co.: easternmost of Grapevine Springs, 30 Jul 2010, M Baker 17145b (ASC)
254	U.S.A.: California. Mono Co.: Fish Slough, 14 Sep 2016, A Howald 3999 (UCR)
270	U.S.A.: Nevada. Lincoln Co.: 27 miles north of Clark County line, 18 Aug 2002, Keil 29916A (OBI)
285	U.S.A.: California. Inyo Co.: Death Valley National Park, Grapevine Springs, north of Lower Vine Ranch west of Scotty’s Ranch Road, 24 Oct 2019, W Hodgson 32480 (DES)
286	U.S.A.: California. Inyo Co.: Death Valley NP, Grapevine Springs, along road to Lower Vine Ranch W of Lower Vine Ranch and W of Scotty's Ranch Road, 24 Oct 2019, W Hodgson 32481 (DES)
288	U.S.A.: California. Inyo Co.: Death Valley National Park, Surprise Spring West (Big Surprise Spring no. 1 in report), east of Scotty's Castle Road, 28 Oct 2019, W Hodgson 32493 (DES)
289	U.S.A.: California. Inyo Co.: Death Valley National Park, Surprise Spring West (Big Surprise Spring no. 2 in report), east of Scotty's Castle Road, 28 Oct 2019, W Hodgson 32494 (DES)
290	U.S.A.: California. Inyo Co.: Death Valley National Park, northeast end of Nevares Springs, east of Park Village, 24 Oct 2019, W Hodgson 32485 (DES)
301	U.S.A.: California. San Bernardino Co.: Mojave Desert: Cushenbury Spring, 31 Aug 2010, J Wood & D Ray 2467 (RSA)
313	U.S.A.: Nevada. Clark Co.: Hidden Valley, 28 Aug 2012, R Johnson 12-063 (BRY)
318	U.S.A.: Nevada. Clark Co.: Warm Springs, Moapa, 17 Jul 2012, R Johnson 12-036 (BRY)
320	U.S.A.: Nevada. Clark Co.: Red Bluff Spring, 21 Jun 1988, D Atwood 13374 (BRY)
330	U.S.A.: California. San Bernardino Co.: San Bernardino Mountains, along the Pacific Crest Trail up a small draw one mile east of Highway 178 trailhead, 26 Apr 2017, G Rink 14578 (ASC)
331	U.S.A.: Arizona. Mohave Co.: Desert springs E of Littlefield, 22 Sep 2016, G Rink 14447 (ASC)
332	U.S.A.: Arizona. Mohave Co.: Tassi Spring, 11 Aug 2011, G Rink 10834 (ASC)
3348*	U.S.A.: Arizona. Coconino Co.: Grand Canyon, Hualapai Nation, near Medicine Spring, just upstream from RM 180, ca 50 m south of river, river left, 07 Jun 2015, W Hodgson 30437 (DES)
338	U.S.A.: Arizona. Coconino Co.: Warm Springs downstream of Lava Falls, 30 Sep 2007, G Rink 6711 (ASC)
339	U.S.A.: Arizona. Mohave Co.: Desert Springs, on the east side of I-15 and the east side of the Virgin River just east of Littlefield, 01 May 2016, G Rink 13779 (ASC)
340	U.S.A.: Utah. Washington Co.: Red hills in the NW part of St. George behind residence at 496 Diagonal Street, 28 Jul 1982, L Higgins 13155 (RENO)

341	U.S.A.: Nevada. Mineral Co.: Sodaville, 30 Aug 1977, M Williams 77-95-2 (RENO)
342	U.S.A.: Arizona. Mohave Co.: Hualapai Indian Reservation, Grand Canyon, Travertine Falls, Colorado River Mile 230.5, 22 Mar 1995, W Hodgson 8514 (DES)
343	U.S.A.: California. San Bernardino Co.: San Bernardino National Forest, Cushenbury Springs, lower end of Cushenbury Grade, 20 Aug 1982, R Thorne 55092 (RSA)
344	U.S.A.: Nevada. Clark Co.: Spring Mountains, seep north of town of Mountain Springs, Hwy 160, 21 Jul 2012, R Johnson 12-052 (BRY)
358	U.S.A.: California. San Bernardino Co.: Box "S" Springs on N side of San Bernardino Mtns below Cushenberry Canyon, 11 Jun 1986, A Sanders 6597 (UCR)
359	U.S.A.: California. San Bernardino Co.: “upper Furnace Spring,” in nameless canyon 2 canyons west of Furnace Canyon, 18 Aug 1998, A Sanders 22208 (UCR)
360	U.S.A.: California. San Bernardino Co.: Furnace Spring Canyon, at spring (Gordon Spring?) 4200 ft. SW of Furnace Spring, 300 ft. south of dirt road, 18 Aug 1998, D Bramlet 2696 (UCR)
361	U.S.A.: Nevada. Clark Co.: At the base of the second group of palms going south along Honey Bee Pond, NE edge of area, 21 Jul 1995, J Alexander 4138 (BRY)
363	U.S.A.: Utah. Washington Co.: Seep along road ca 0.5 mi east of Danish Ranch, 19 Aug 1986, R Warrick 2961 (BRY)
364	U.S.A.: Arizona. Coconino Co.: Grand Canyon, Hualapai Reservation, Three Springs, river left, 24 Sep 2009, W Hodgson 24461 (DES)
365	U.S.A.: Arizona. Coconino Co.: Grand Canyon, Hualapai Nation, near Medicine Spring, just upstream from RM 180, ca 50 m south of river, river left, 07 Jun 2015, W Hodgson 30437 (DES)
367	U.S.A.: Nevada. Nye Co.: Ash Meadows NWR, Amargosa Desert, at intersection of Point of Rocks Rd. and the main Ash Meadows County Road, 28 Jun 1996, J Alexander 569 (DES)
368	U.S.A.: Arizona Mohave Co.: Hualapai Indian Reservation, Grand Canyon, up Bridge Canyon within 1/2 mile south of Colorado River at RM 235, 11 Sep 1994, W Hodgson 8517 (DES)
369	U.S.A.: Arizona. Coconino Co.: Grand Canyon, Hualapai Indian Reservation, Three Springs, lower end near trail and creek, 09 Aug 2011, W Hodgson 26526 (ASU)
370	U.S.A.: Arizona. La Paz Co.: Palmerita Ranch 7.5. W-most spring of Grapevine Springs; S. side of Santa Maria River, 23 Mar 1980, Butterwick 5851 (ASU)
371	U.S.A.: Arizona. Mohave Co.: Virgin River near Littlefield, 04 Aug 1986, R Gierisch 4905 (ASU)
372	U.S.A.: California. Inyo Co.: Last Chance Range, Eureka Valley drainage: East of north end of Eureka Sand Dunes, 10 May 1977, M DeDecker 5234 (RSA)
373	U.S.A.: California. Inyo. Co.: Alabama Hills, west of Lone Pine: Junction of Whitney Portal and Horseshoe Meadows Road, 21 Jul 1982, V Yoder 5378 (RSA)
376	U.S.A.: Nevada. Lincoln Co.: Pahranagat National Wildlife Refuge, south end of upper Pahranagat Lake, 29 Aug 1980, A Tiehm 6275 (RENO)
377	U.S.A.: Nevada. Lander Co.: Big Smoky Valley, Spencer Hot Spring at the NE end of the valley, 12 Jul 1984, A Tiehm 8924 (BRY)
378	U.S.A.: Nevada. Nye Co.: Duckwater Valley, Big Warm Springs near Duckwater, 15 Sep 1983, A Tiehm 8409 (BRY)

379	U.S.A.: Arizona. Mohave Co.: East of Littlefield and Virgin River, west of Farm Rd, south of Little Jamaica Natural Swimming Hole south of I-15, 19 Jul 2019, W Hodgson 32282 (DES)
380	U.S.A.: Arizona. Mohave Co.: Grand Canyon, Hualapai Indian Reservation, Three Springs Canyon, 17 May 2006, W Hodgson 21178 (DES)
381	U.S.A.: Arizona. Mohave Co.: Grand Canyon, Hualapai Indian Reservation, Travertine Canyon near Colorado River, RM 229 RL, 24 Feb 2010, W Hodgson 24533 (DES)
382	U.S.A.: Arizona. Mohave Co.: Grand Canyon, Cave Canyon, just below Columbine/Emory Falls, river left, 01 Nov 2020, W Hodgson 32711 (DES)
384	U.S.A.: California. San Bernardino Co.: Rabbit Springs, one mile NW of Lucerne Valley town center; Mojave Desert, 07 Jun 1978, F Vasek s.n. (UCR)
385	U.S.A.: California: San Bernardino Co.: Near Cushenbury Springs, c. 0.25 miles north of 18 and Camp Rock Rd., Mojave Desert/San Bernardino Mts, 13 Aug 1998, M Provance 1012 (UCR)
386	U.S.A.: California. Mono Co.: Fish Slough: 4.5 mi. N of Bishop at northern end of Owens Valley. Along channel below Rocked-In Spring, NE end of slough, 14 Jul 1984, W Ferren 419 (RSA)
387	U.S.A.: Arizona. Mohave Co.: Grand Canyon, Hualapai Nation Reservation, Bridge Canyon, RM 235, river left, 25 Feb 2010, W Hodgson 24546 (RSA)
388	U.S.A.: California. San Bernardino Co.: Transverse Ranges; San Bernardino Mountains region: Lone Valley; SE of Arrastre Creek in spring area near ranch, 06 Jun 2014, N Fraga 4869 (RSA)
389	U.S.A.: California. San Bernardino Co.: Bighorn Mountains; Arrastre Creek in narrows below Horsethief Flat about a 1/2 mile SE of Terrace Springs, 09 Jul 2014, D Bell 7376 (RSA)
393	U.S.A.: California. San Bernardino Co.: Upper Box Spring, 0.35 mi. below Camp Rock Rd., San Bernardino Mtns./Mojave Desert, 22 Sep 1996, A Sanders 19563 (UCR)
394	U.S.A.: California. Inyo Co.: Toll House (Batchelder) Spring, on State Hwy 168, c. 5 miles west of Westgard Pass, Inyo National Forest; White Mountains, 05 Jul 1981, R Goeden s.n. (UCR)
395	U.S.A.: Nevada. Clark Co.: 5 mi. NW of Las Vegas on highway to Beatty at or near Artesian Well, 16 Jun 1937, I LaRivers 444 (RENO)
396	U.S.A.: Nevada. Mineral Co.: Near Rock House Springs, S of Marietta, 09 Jul 1980, M Williams 80-195-5 (RENO)
399	U.S.A.: Arizona. Mohave Co.: Grand Canyon, Hualapai Nation, Travertine Canyon, just below Travertine Falls. 30 Oct 2020, W Hodgson 32684 (DES)
401	U.S.A.: Nevada. Nye Co.: Saline seeps just S of Cold Springs at edge of large alkali playas 4 air miles SW of Mt. Annie, 17 Jun 1980, A Tiehm 5949 (RENO)
404	U.S.A.: California. San Bernardino Co.: Near Box Springs, near base of Cushenbury Grade; N side of San Bernardino Mtns, 27 May 1967, O Clarke s.n. (UCR)
4077*	U.S.A.: Arizona. Coconino Co.: Grand Canyon, Three Springs Canyon, 16 May 2011, W Hodgson 26148 (DES)