Article Type: Special Issue Article

RESEARCH ARTICLE

INVITED SPECIAL ARTICLE

For the Special Issue: Using and Navigating the Plant Tree of Life

Short Title: Walker et al.—Phylotranscriptomic analysis of Caryophyllales

From cacti to carnivores: Improved phylotranscriptomic sampling and hierarchical homology inference provide further insight into the evolution of Caryophyllales

Joseph F. Walker[1,13], Ya Yang[2], Tao Feng[3], Alfonso Timoneda[3], Jessica Mikenas[4,5], Vera Hutchison[4], Caroline Edwards[4], Ning Wang[1], Sonia Ahluwalia[1], Julia Olivieri[4,6], Nathanael Walker-Hale[7], Lucas C. Majure[8], Raúl Puente[8], Gudrun Kadereit[9,10], Maximilian Lauterbach[9,10], Urs Eggli[11], Hilda Flores-Olvera[12], Helga Ochoterena[12], Samuel F. Brockington[3], Michael J. Moore,[4] and Stephen A. Smith[1,13]

Manuscript received 13 October 2017; revision accepted 4 January 2018.

[1] Department of Ecology & Evolutionary Biology, University of Michigan, 830 North University Avenue, Ann Arbor, MI 48109-1048 USA

[2] Department of Plant and Microbial Biology, University of Minnesota-Twin Cities, 1445 Gortner Avenue, St. Paul, MN 55108 USA

[3] Department of Plant Sciences, University of Cambridge, Cambridge CB2 3EA, UK

[4] Department of Biology, Oberlin College, Science Center K111, 119 Woodland Street, Oberlin, OH 44074-1097 USA

[5] Current address: USGS Canyonlands Research Station, Southwest Biological Science Center, 2290 S West Resource Blvd, Moab, UT 84532 USA

[6] Institute of Computational and Mathematical Engineering (ICME), Stanford University, 475 Via Ortega, Suite B060, Stanford, CA, 94305-4042 USA

[7] School of Biological Sciences, Victoria University of Wellington, Kelburn Parade, Kelburn, Wellington, 6012, New Zealand

[8] Department of Research, Conservation and Collections, Desert Botanical Garden, 1201 N. Galvin Pkwy, Phoenix, AZ 85008 USA

[9] Institut für Molekulare Physiologie, Johannes Gutenberg-Universität Mainz, D-55099 Mainz, Germany

[10] Institut für Molekulare und Organismische Evolutionsbiologie, Johannes Gutenberg-Universität Mainz, D-55099 Mainz, Germany

[11] Sukkulenten-Sammlung Zürich / Grün Stadt Zürich, Mythenquai 88, CH-8002 Zürich, Switzerland

[12] Departamento de Botánica, Universidad Nacional Autónoma de México, Apartado, Postal 70-367, 04510, Mexico City, Mexico

[13] Authors for correspondence (e-mails: jfwalker@umich.edu (Walker) and eebsmith@umich.edu (Smith))

Citation: Walker, J. F., Y. Yang, T. Feng, A. Timoneda, J. Mikenas, V. Hutchison, C. Edwards, et al. 2018. From cacti to carnivores: Improved phylotranscriptomic sampling and hierarchical homology inference provide further insight into the evolution of Caryophyllales. *American Journal of Botany* 105(3): XXX.

DOI: XXXX

PREMISE OF THE STUDY: The Caryophyllales contains ~12,500 species and is known for its cosmopolitan distribution, convergence of trait evolution, and extreme adaptations. Some relationships within the Caryophyllales, like those of many large plant clades, remain unclear, and phylogenetic studies often recover alternative hypotheses. We explore the utility of broad and dense transcriptome sampling across the order for resolving evolutionary relationships in Caryophyllales.

METHODS: We generated 84 transcriptomes and combined these with 224 publicly available transcriptomes to perform a phylogenomic analysis of Caryophyllales. To overcome the computational challenge of ortholog detection in such a large data set, we developed an approach for clustering gene families that allowed us to analyze >300 transcriptomes and genomes. We then inferred the species relationships using multiple methods and performed gene-tree conflict analyses.

KEY RESULTS: Our phylogenetic analyses resolved many clades with strong support, but also showed significant gene-tree discordance. This discordance is not only a common feature of phylogenomic studies, but also represents an opportunity to understand processes that have structured phylogenies. We also found taxon sampling influences species-tree inference, highlighting the importance of more focused studies with additional taxon sampling.

CONCLUSIONS: Transcriptomes are useful both for species-tree inference and for uncovering evolutionary complexity within lineages. Through analyses of gene-tree conflict and multiple methods of species-tree inference, we demonstrate that phylogenomic data can provide unparalleled insight into the evolutionary history of Caryophyllales. We also discuss a method for overcoming computational challenges associated with homolog clustering in large data sets.

KEY WORDS: Agdestidaceae; Amaranthaceae; Caryophyllales; coalescent; gene-tree conflict; homology; phylogenomics; phylotranscriptomic; supermatrix

The Caryophyllales (sensu Angiosperm Phylogeny Group IV [APG IV, 2016]) contain an estimated ~12,500 species and are found on all continents and in all major terrestrial ecosystems (Hernández-Ledesma et al., 2015). The clade is notable not only for its diversity and broad ecological and geographic distribution, but also for its array of unique morphological and ecophysiological adaptations. Many Caryophyllales (most famously, many cacti) are noted for their extreme drought tolerance, but the clade also contains species that exhibit extreme cold tolerance (Cavieres et al., 2016), halophytism (Flowers and Colmer, 2008; White et al., 2017), heavy metal hyper-accumulation (Moray et al., 2016), carnivory (e.g. Venus flytrap, sundews, and *Nepenthes* pitcher plants) (Albert et al., 1992; Givnish, 2015), betalain pigmentation (Brockington et al., 2015), C_4 and CAM photosynthesis (Wang et al., unpublished manuscript; Sage et al., 2011; Moore et al., 2017; Sage, 2017), and succulence (Sajeva and Mauseth, 1991; Eggli and Nyffeler, 2009). Most of these adaptations are known to have arisen multiple times throughout the clade, making Caryophyllales a key natural laboratory for understanding trait evolution in angiosperms. The clade also includes numerous economically important species (e.g., beets, quinoa, and spinach), bolstering its utility as a model system for understanding morphological and physiological evolution.

Previous phylogenetic work, focused on resolving the backbone relationships of Caryophyllales, has utilized morphology (Rodman et al., 1984), targeted gene sequencing (Rettig et al., 1992; Brockington et al., 2009, 2011; Schäferhoff et al., 2009), plastome sequencing (Arakaki et al., 2011), and transcriptome data (Yang et al., 2015, 2018). These studies have resulted in the expansion of the traditional Caryophyllales (i.e., corresponding essentially with the original Centrospermae) to include other families (e.g., Polygonaceae, Plumbaginaceae, Droseraceae, Rhabdodendraceae) and the recircumscription of a number of families, especially the division of previously broadly circumscribed Molluginaceae, Phytolaccaceae, and Portulacaceae (APG IV, 2016). These taxonomic rearrangements have resulted in the 38 families currently recognized by APG IV (2016) as well as the more recently proposed Corbichoniaceae (Thulin et al., 2016). Almost all of these families have been shown to be monophyletic, with the possible exception of Phytolaccaceae due to the uncertain position of the tropical liana *Agdestis clematidea* (Hernández-Ledesma et al., 2015). Our understanding of relationships among these families has advanced greatly during the past 20 years. For example, there has been consistent support at the base of the extant Caryophyllales for a split between the noncore Caryophyllales, consisting of the carnivorous families (Droseraceae, Drosophyllaceae, Nepenthaceae, Ancistrocladaceae, Dioncophyllaceae) and allies (Tamaricaceae, Frankeniaceae, Polygonaceae, and Plumbaginaceae), and a larger clade containing the remaining diversity of the order (Brockington et al., 2009; Hernández-Ledesma et al., 2015). Within the latter clade, there is support for a grade composed of four species-poor families (Rhabdodendraceae, Simmondsiaceae, Asteropeiaceae, and Physenaceae) that leads to a well-supported clade containing all of the core members of Caryophyllales (i.e., the old Centrospermae) (Hernández-Ledesma et al., 2015). The diversification within several clades was apparently very rapid (Arakaki et al., 2011), making resolution of the backbone phylogeny of this clade difficult (Hernández-Ledesma et al., 2015). The use of genomic data (Jarvis et al., 2014; Fontaine et al., 2015), restriction-site-associated DNA sequencing (RADSeq) (Eaton et al., 2016), genotyping by sequencing (Fernández-Mazuecos et al., 2017), and transcriptome data (Dunn et al., 2008; Smith et al., 2011; Cannon et al., 2015; Pease et al., 2016) have all proven to be robust tools for inferring recalcitrant evolutionary relationships at both shallow and deep time scales, but to date these tools have not been applied to Caryophyllales with sufficient taxon sampling to test hypotheses of early-diverging relationships.

Transcriptomes hold considerable promise as a phylogenetic tool as they provide a relatively cost-effective way to generate a wealth of sequence data for evolutionary analyses, including the exploration of gene-tree conflict and gene/genome duplications (Wickett et al., 2014; Cannon et al., 2015; Smith et al., 2015; Yang et al., 2018). For example, in a study using 92 transcriptomes to reconstruct land-plant relationships, Wickett et al. (2014) demonstrated that phylotranscriptomic data sets provide highly informative data for resolving deeper-level phylogenetic relationships, but some relationships were sensitive to the reconstruction method. Underlying these sensitive relationships is often gene-tree conflict that may arise from a variety of biological causes, including but not limited to incomplete lineage sorting (ILS), hybridization, hidden paralogy, and horizontal gene transfer (Galtier and Daubin, 2008). Gene-tree conflict makes it difficult to assess species relationships, as phylogenetic hypotheses are the product of the genes selected for an analysis (Maddison, 1997; Rokas et al., 2003; Walker et al., 2014; Smith et al., 2015), and individual genes can have overwhelming influences on the species-tree topology in phylogenomic data sets (Brown and Thomson, 2017; Shen et al., 2017; Walker et al., 2018). Large multi-locus data matrices may also result in artificially inflated support (Seo, 2008), masking underlying conflict. Furthermore, taxon sampling can affect phylogenetic reconstruction using both coalescent and supermatrix methods (Wickett et al., 2014; Walker et al., 2017). These problems, however, are not a consequence of using transcriptomes per se—rather, transcriptome analyses have exposed problems that have always been present but have been overlooked due to limited data sets. In short, the use of transcriptome data sets provides novel insights into evolutionary history and leads to biological insights that are not obtainable from a handful of loci (Yang et al., 2015; Pease et al., 2016; Smith et al., 2017).

We explore the conflict underlying relationships across the phylogenetic backbone of Caryophyllales using a data set consisting of 295 transcriptomes and three genomes, collectively comprising 32 of the 39 families of Caryophyllales (APG IV, 2016; Thulin et al., 2016). Due to the severe computational burden imposed by the exponential scaling of all-by-all BLAST during homolog detection, we outline a method of homolog clustering through post-order traversal (tip-to-root). This process allowed us to conduct the all-by-all procedure on individual clades that are then combined in a hierarchical manner. Our analyses highlight the tremendous power of using large data sets for inferring species relationships, but they also reveal some of the limitations of large phylogenomic analyses for species relationship inference.

<h1>MATERIALS AND METHODS

<h2>Data availability

The raw reads for transcriptomes generated for this study have been deposited in the NCBI sequence read archive (Bioproject SRP127816). Assemblies, orthologous gene clusters, alignments, and trees are available in the Dryad Digital Repository (https://datadryad.org//review?doi=doi:10.5061/dryad.470pd). Scripts and programs written for this project can be found at Bitbucket (https://bitbucket.org/jfwalker/ajb_bigtree).

<h2>Taxon sampling, tissue collection, sequencing, and read assembly

Taxon sampling was designed to broadly cover Caryophyllales. In total, our sampling includes 295 Caryophyllales transcriptomes and three Caryophyllales genomes, representing 298 species and 32 of the 39 families in the clade; the phylogenetic distribution of the species sampled is shown in a collapsed genus-level tree of Smith et al. (2018) (Fig. 1). The families Asteropeiaceae, Barbeuiaceae, Corbichoniaceae, Dioncophyllaceae, Halophytaceae, Lophiocarpaceae, and Rhabdodendraceae were not sampled due to the difficulty of obtaining fresh tissue of these taxa. We also included *Agdestis clematidea* to test its phylogenetic position within the phytolaccoid clade (Nyctaginaceae, Petiveriaceae, Phytolaccaceae s.l., Sarcobataceae). We sampled 10 outgroups spanning the asterids (*Mimulus guttatus*, *Solanum lycopersicum*, *Ilex paraguariensis*, *Actinidia deliciosa*, *Vaccinium corymbosum*, *Camptotheca acuminata*, and *Davidia involocrata*), rosids (*Vitis vinifera*), Ranunculales (*Aquilegia coerulea*), and Santalales (*Taxillus nigrans*). A summary of all 84 newly generated transcriptomes can be found in Appendix S1 (see Supplemental Data with this article), along with the sources of the data for previously generated transcriptomes (Appendix S2). In many cases, a previously assembled transcriptome was used, in which case we listed the digital object identifier (doi) for the Dryad Digital Repository where that assembly was downloaded.

The 84 newly generated transcriptomes were sequenced and processed following the previously developed phylotranscriptomic workflow (Yang et al., 2017). In short, RNA was obtained from fresh tissue that was flash frozen in liquid nitrogen and stored at −80°C. When possible, the RNA was extracted from a mixture of both young leaf and flower bud. The various methods used for the newly generated transcriptomes can be found in Appendix S1. All RNA sequencing (RNA-seq) libraries were stranded to simplify assembly and translation. Paired-end sequencing for all newly generated transcriptomes was performed using Illumina HiSeq platforms (see Appendix S1 for additional details). Sequence assembly and translation were conducted using previously designed protocols as outlined in Brockington et al., 2015; any differences are highlighted in Appendix S1.

<h2>Construction of species trees

We conducted two analyses to reconstruct the relationships within the Caryophyllales. In the first, we conducted a hierarchical clustering method across the entire Caryophyllales (abbreviated ALL throughout), and in the second, we conducted targeted analyses on each well-sampled major group (abbreviated IND throughout).

<h2>Reconstruction of the Caryophyllales species tree with hierarchical clustering (ALL)

<h3>Tip clustering

The code developed and used for this project can be found at (https://bitbucket.org/jfwalker/ajb_bigtree) and the overarching procedure of tip-to-root clustering has been incorporated into PyPHLAWD (Smith and Brown, 2018, in this issue). This method utilizes a taxonomic tree based on previous phylogenetic hypotheses. Homologs were first clustered by binning transcriptomes within taxonomic families (which we refer to as tip clustering), and clustering then worked backward toward the root of a taxonomic tree (internal node clustering; Fig. 2.). Hence, taxonomic families were the tips for the post-order clustering (Fig. 2). The family Amaranthaceae was separated into Chenopodiaceae and Amaranthaceae (Hernández-Ledesma et al., 2015). *Agdestis clematidea* was treated as its own tip within the monotypic Agdestidaceae (see Results) and was clustered with the monotypic families due to its conflicting phylogenetic positions (Hernández-Ledesma et al., 2015). The addition of these two families into the analysis expands the total Caryophyllales sampling to 34 families and the total possible families for Caryophyllales to 41 (i.e., the 38 families recognized by APG IV [2016], plus Corbichoniaceae, Agdestidaceae, and Chenopodiaceae). The analysis was conducted on 19 bins of families with three or more species represented, 1 bin for all families with less than three species represented, and one bin for the outgroups, for a total of 21 bins (Fig. 2). The size of these bins ranged from as many as 39 individual species in Caryophyllaceae and Chenopodiaceae, to as few as three in a variety of families (Fig. 2). The first step for all transcriptomes and genomes was to reduce sequence redundancy in the translated amino acid data sets using cd-hit (-c 0.995 –n 5) (Fu et al., 2012). Clades including three taxa or more were clustered into putative homolog groups following (Yang and Smith, 2014); The method consists of conducting an all-by-all BLASTP (Altschul et al., 1997), with an E-value cutoff of 10. The top 1000 hits were retrieved and putative homolog groups were retained for species clusters with a hit fraction >0.4. Subsequently, Markov clustering was conducted as implemented in mcl (Van Dongen, 2000), with the inflation value cutoff set to 1.4 and the E-value cutoff set to 10^{-5} "–abc –te 18 –tf 'gq(5)' ". Resulting phylogenetically informative clusters (≥4 sequences) were separated out for further filtering and remaining clusters (<4 sequences) being retained for node-level clustering.

Clusters with four or more sequences were then aligned using MAFFT v7 (Katoh and Standley, 2013), conducted for 1000 cycles of iterative refinement, with the setting "–auto –amino –maxiterate 1000". The alignments were then trimmed for 10% column occupancy using the phyx (v.0.99) program pxclsq (Brown et al., 2017) with the settings "–p 0.1 –a". After each approximation, roughly 10 homolog clusters were manually checked to ensure the alignment and cleaning procedures were performed properly. Phylogenetic trees were then estimated for each potential homolog cluster through maximum likelihood using the RaxML v8.2.3 (Stamatakis, 2014) algorithm (for <100 sequences) and the FastTree2 v2.1.8 (Price et al., 2010) algorithm (for >100 sequences). In both cases, the trees were estimated under the WAG model of protein evolution.

Each inferred homolog cluster then had all putatively spurious tips filtered out by removing tips based on relative and absolute branch length criteria outlined by Yang and Smith (2014). The absolute tip cutoff used was three substitutions per site, and the relative tip cutoff was two substitutions per site. These values were used because anything of that length or greater likely represented poor alignment or some form of long-branch attraction based upon conserved domain regions and could lead to compounding issues in downstream alignment and tree inference. The homolog tree was analyzed for all clades that consisted solely of genes from the same taxa; these were then condensed down to a single tip, which was chosen based on the criterion of having the most potentially informative sites (i.e., most amino acids in trimmed alignment). The condensing of these clades was carried out, because any clade consisting solely of tips from the same individual was likely the product of different isoforms or in-paralogs, neither of which provides a means of inferring species relationships. The sequences of the remaining tips were then extracted to form new homolog clusters, with which the same process was again performed two more times for further refinement. The bin containing all small families and the bin containing all outgroups were separately combined and clustered using the same method as the individual families.

<h3>Internal node clustering

Clustering at internal nodes of the taxonomy tree was conducted using a post-order tree traversal method (tip-to-root), which was performed following the predicted topology from the Angiosperm Phylogeny Website (Stevens, 2001) (Fig. 2), which itself represents a continuously updated compilation of previously inferred phylogenies (e.g., Cuenod et al., 2002; Brockington et al., 2009; Christenhusz et al., 2014). The method proceeds by using the pre-clustered groups generated by the tip clustering step. The predicted sister tips are the first to be combined (e.g., Cactaceae and Portulacaceae; Fig. 2). The combination occurs by first creating a BLAST database from one of the tips (or a node depending on where the clustering is occurring). This database consists of random representatives from each of the clustered homologous genes. The number of random representatives was determined by the size of the homologous gene cluster. For clusters with fewer than four sequences, all sequences from the cluster were used; for clusters with four or more sequences, 4 + sqrt (no. of sequences in the cluster) were randomly selected and added to the database to allow for proportional representation of the cluster.

After the database was initiated from one tip, a BLAST analysis was performed for representative sequences from the other tip, with the representatives being chosen based on the same criteria. The BLASTP analysis was conducted using an e-value cutoff of 1e-3, and only the top hit was retrieved. All clusters from one sister tip/node were then combined with their top hit from the other sister tip/node, using a one-sided BLAST approach. If multiple hits occurred between the two, then the new node cluster was formed consisting of all homologous gene families that had a hit.

For example, in the case of Portulacaceae and Cactaceae, the new node level cluster "Cactaceae+Portulacaceae" theoretically could contain all 44 representative taxa from those two families. The next step for the inferred homologs at the node level "Cactaceae+Portulacaceae" bifurcate is to combine Anacampserotaceae, with the newly formed homolog cluster of (Cactaceae+Portulaceae) labeled "1" on Fig. 2, which in turn would form the cluster Cactaceae+Portulaceae+Anacampserotaceae, labeled "6" on Fig. 2. In later steps, only clusters with less than 5000 sequences are retained as future tree building, and alignment steps often have issues with such large data sets (Fig. 2).

Although predominantly conducted in a post-order means, or from tip-to-root, the procedure included some deviations. After that the cluster containing the noncore Caryophyllales, single families, and outgroups was then combined with the core Caryophyllales (internal node 20, Fig. 2). This method results in a significant decrease in computational burden imposed by large homolog groups, but due to the removal of clusters smaller than 5000 sequences also causes the final homolog clusters to be smaller than those usually produced by an all-by-all BLAST.

<h3>Inference of final gene trees

After the formation of homolog clusters, inference of the final gene trees was conducted by first aligning with MAFFT and trimming the aligned matrix with pxclsq with the settings described above. In the first round of gene-tree inference, FastTree2 v2.1.8 was used with the same settings as noted above to infer all individual gene trees. Next, all sequences with an absolute branch length of 2 substitutions/site and a relative branch length of 1 substitution/site were trimmed. Furthermore, any clades that consisted of only genes from a single taxon were again trimmed down to only the gene with the highest number of aligned characters. Next, any clade including genes from at least four taxa as well as a branch with at least one substitution/site was split into a separate homolog group. The same process was then repeated to help further refine the data set.

Orthologous sequences were inferred from the inferred homologous gene trees using the rooted tree (RT) method (Yang and Smith, 2014) and specifying *Aquilegia caerulea* (Ranunculaceae), *Taxillus nigrans* (Loranthaceae), and *Vitis vinifera* (Vitaceae) as outgroups with a minimum of 50 sequences required as ingroup taxa. The specification of three outgroups in the RT method meant that the final tree contained 305 of the 308 taxa used in the analysis. The other outgroup taxa were kept as ingroups as they are predicted to form a clade with Caryophyllales and needed to root the species tree after final inference. The orthologous genes were then aligned using MAFFT with the settings above, cleaned with pxclsq for 30% column occupancy, and only alignments that still contained at least 150 characters were retained after cleaning. Gene trees were then estimated using RAxML, tips longer than 0.8 substitutions/site were removed, and any internal branches longer than 0.8 substitutions/site or greater were separated. Then clades with fewer than 50 sequences or fewer than 17 different families were removed from the species tree analysis. The resulting set of 1238 gene trees was used for the downstream MQSST species tree analysis. Finally, we filtered for genes that contained at least 17 different families and 200 taxa, resulting in 58 orthologs for the supermatrix analysis.

<h3>Species-tree inference for Caryophyllales (ALL)

We inferred a species tree for the data set including all of the Caryophyllales with two methods. First, we conducted a maximum likelihood analysis, as implemented in RAxML v8.2.3, on a supermatrix of 58 orthologs concatenated using pxcat (from phyx; Brown et al., 2017). The supermatrix was partitioned by ortholog, with the WAG substitution model specified for each partition; final inference was conducted using Γ rate variation (PROTCATWAG in RAxML). Support for the tree was evaluated by running 100 rapid bootstraps as implemented in RAxML and for 200 replicates of the quartet sampling method (Pease et al., 2018). The second method employed the Maximum Quartet Support Species Tree (MQSST) algorithm as implemented in ASTRAL-II (v.4.10.12) (Mirarab and Warnow, 2015). This was conducted using the 1238 orthologs that contained at least 17 families and 50 taxa. Support for the tree was inferred using local posterior probabilities (Sayyari and Mirarab, 2016).

<h2>Reconstruction of densely sampled clades within Caryophyllales with individual analyses (IND)

Although our hierarchical clustering method was effective in overcoming the computational challenge in orthology inference, taxon sampling may affect orthology inference due to a variety of reasons (e.g., heterogeneity in evolutionary rates, gene/genome duplication, etc.). As such, we also conducted species-tree analyses on the five individual clades of interest to help verify the species relationships obtained from using the 305 taxa data set. The densely sampled clades we chose to analyze separately included Nyctaginaceae, Caryophyllaceae, Amaranthaceae+Chenopodiaceae, Cactaceae, and the clade of noncore Caryophyllales. The methods and settings used for tree inference in each case varied, given the heterogeneity in evolutionary rates across each of the separate clades; therefore, we have outlined settings and modifications below. All statistics, as reported by pxlssq (from phyx; Brown et al., 2017) for the final matrices, can be found in Appendix S3.

<h3>Caryophyllaceae

Clustering was performed using the same methods as the tip level clustering, but included three Chenopodiaceae (*Spinacia oleracea*, *Chenopodium quinoa*, and *Beta vulgaris*), two Amaranthaceae (*Alternanthera brasiliana* and *Tidestromia lanuginosa*), and *Achatocarpus gracilis* (Achatocarpaceae) as outgroups. The homolog trees then had spurious tips trimmed using an absolute cutoff of two substitutions/site, and the monophyletic tips were then masked leaving the tip with the most aligned characters. Orthologs were identified using maximum inclusion (MI) (Yang and Smith, 2014). Of these identified orthologs, groups containing at least 40 of the 45 taxa were chosen, resulting in 999 inferred orthologs. The individual orthologs were then aligned with MAFFTv7, with the settings "--auto --amino --maxiterate 1000", and alignment trimmed for 10% minimum occupancy using pxclsq (–p 0.1 –a), and a ML tree was inferred using RAxML v8.2.3 for each ortholog.

The species tree was inferred using the same two methods as above (i.e., using MQSST as implemented in ASTRAL-II and through a supermatrix ML analysis using FastTree to generate an input topology for a more thorough analysis using RAxML v.8.2.3. In both ML analyses, the WAG model of evolution was used, with partitioning by gene to ensure that a separate rate was estimated for each gene using CAT.

<h2>Nyctaginaceae

The node level clustering that contained the Nyctaginaceae (37 taxa) and Petiveriaceae (4 taxa) was used for inference of the clade (node 2; Fig. 2). Initially, homolog groups, which were found to contain at least 1000 genes and sequences from both Nyctaginaceae and Petiveriaceae, were aligned using MAFFTv.7, cleaned with pxclsq for 10% matrix occupancy, and homolog trees were inferred with FastTree v2.1.8 under the WAG model of evolution. Next, spurious tips with a relative value of 1 substitution/site and an absolute value of 2 substitutions/site were removed, and monophyletic tips were masked conserving the tip with the highest number of aligned characters. Next, orthologs were inferred using the maximum inclusion procedure, searching for ortholog groups containing at least 40 of the 41 taxa, which resulted in 389 orthologs for the analysis. Species trees were inferred using the method mentioned above for Caryophyllaceae.

<h3>Cactaceae

The species trees were inferred using the same method as Nyctaginaceae with the following minor modifications. The cluster used was the node-level cluster that consisted of Cactaceae (29 taxa), Portulacaceae (8 taxa), and Anacampserotaceae (3 taxa) (node 6; Fig. 2). The ortholog groups were filtered for those consisting of at least 40 of the 47 taxa, which resulted in 1502 orthologs.

<h3>Amaranthaceae and Chenopodiaceae

The species trees were inferred using the same method as Nyctaginaceae with these minor modifications: we used homologous gene clusters of 500 sequences or fewer, as opposed to 1000. The clusters used were from the node-level cluster that consisted of Amaranthaceae (21 taxa), Chenopodiaceae (39 taxa), and five representative Caryophyllaceae (node 8; Fig. 2). The ortholog groups were filtered for those consisting of at least 60 of the 65 taxa, which resulted in 455 orthologs.

<h3>The noncore Caryophyllales

The species trees were inferred using the same method as Nyctaginaceae with the following modifications. The cluster used was the node-level cluster consisting of Polygonaceae (37 taxa), Plumbaginaceae (4), Tamaricaceae (3), Nepenthaceae (3), and Droseraceae (4) (node 12; Fig. 2). The node 12 cluster was combined with the clustering of smaller families to add in the other noncore families Drosophyllaceae (1), Ancistrocladaceae (1), and Frankeniaceae (2). The families Basellaceae (2), Microteaceae (1), Physenaceae (1), and Simmondsiaceae (1) were added as outgroups. The ortholog groups were filtered for those consisting of at least 55 of the 60 taxa, which resulted in 514 orthologs, of which only 513 contained at least one outgroup and were rooted for the conflict analysis. The final statistics for the supermatrix can be found in Appendix S3.

<h2>Analysis of conflict

We conducted conflict analyses on the trees resulting from the IND analyses using the bipartition-based method as implemented in phyparts (Smith et al., 2015). All gene trees from the clade-specific analyses were rooted by outgroups in a ranked fashion using pxrr (from the phyx package; Brown et al. 2017), whereby, if a taxon in the outgroup is not found, the program searches for the next taxon, thus not requiring all outgroup taxa for rooting. The results were summarized and mapped onto a tree using phypartspiecharts.py (https://github.com/mossmatters/MJPythonNotebooks). A comparison of conflict between the topology of the MQSST and the ML analysis was conducted using pxbp (from the phyx package; Brown et al. 2017), where both trees were rooted on all outgroups using pxrr and the MQSST tree was mapped onto the ML tree.

<h1>RESULTS

We define the support on the MQSST species tree from here on as follows: strong support will correspond to local posterior probabilities (LPP) ≥ 0.95, moderate support will correspond to 0.95 > LPP ≥ 0.80, and low support will correspond to 0.80 > LPP. For the bootstrap (BS) support on the ML tree (Appendix S4), we will consider strong support to be BS ≥90, moderate support will be 90 > BS ≥ 70, and poor support will be anything with BS support lower than 70. Here we also discuss the quartet differential (QD), which reflects the number of alternate topologies a quartet recovers, and the full results of the analysis can be found in Appendix S5. This method provides a means of disentangling a rogue node from one with two dominant topologies, and a thorough description of this form of support and other quartet based support metrics is outlined by Pease et al. (2018, in this issue).

We inferred species relationships using multiple data sets—one data set comprised all taxa (ALL; Fig. 3), whereas the other data sets (described in Appendix S3) included only orthologs inferred from five most densely sampled clades (IND; Figs. 4–8).

<h2>Relationship among major clades across the backbone of Caryophyllales using the ALL data set

Conflict between the MQSST- and ML-inferred phylogenetic trees is shown in Appendix S6. Both analyses recovered a clade of Tamaricaceae+Frankeniaceae sister to Plumbaginaceae+Polygonaceae, which we will collectively refer to as the noncarnivorous noncore (NCNC) clade (Fig. 3; Appendices S4–6). The MQSST analysis had insufficient data to resolve the divergence of the NCNC, resulting in no branch length at the divergence of the carnivorous clade (the families Droseraceae, Ancistrocladaceae, Drosophyllaceae, and Nepenthaceae), and the core Caryophyllales (all other families). In the ML analysis, we recovered the carnivorous clade to be sister to the core Caryophyllales and the NCNC with low support from the ML support statistics (Appendices S4, S5). The majority of nodes within core Caryophyllales received medium to high support in the MQSST and ML trees with notable examples of low support occurring in Amaranthaceae subfamily Polycnemoideae (*Polycnemum majus* and *Nitrophila occidentalis*) and the placement of Cactaceae genera *Leuenbergeria* and *Pereskia*.

The core Caryophyllales was inferred to be nested within a grade of species-poor families (Fig. 3). In the MQSST tree, this grade consisted of Simmondsiaceae, Physenaceae, Microteaceae, and a clade of Stegnospermataceae+Macarthuriaceae diverging in that respective order. In the ML analysis, Limeaceae is nested within the grade, diverging prior to Stegnospermataceae+Macarthuriaceae (Appendices S4–S6). The grade is strongly supported in the MQSST analysis, whereas in the ML analysis there is low bootstrap support for the position of Limeaceae, which in combination with a QD of 0.38 toward a different topology indicates it may have bias toward an alternate position than that recovered by the ML analysis.

Caryophyllaceae was inferred in both analyses to be sister to Achatocarpaceae+Amaranthaceae+Chenopodiaceae, with Amaranthaceae+Chenopodiaceae forming a clade sister to Achatocarpaceae. In the MQSST analysis, Chenopodiaceae was monophyletic; however, the subfamily Polycnemoideae was not nested within Amaranthaceae, making Amaranthaceae paraphyletic without the inclusion of Chenopodiaceae. In the ML analysis, there was low support for a clade consisting solely of genus *Beta* and Polycnemoideae, making Chenopodiaceae paraphyletic without Amaranthaceae. Sister to the clade containing Amaranthaceae and Chenopodiaceae, the family Achatocarpaceae was recovered as monophyletic with strong support by both BS and LPP, and no common discordant topologies were found from the QS analysis.

Sister to the clade of Amaranthaceae and relatives was a clade encompassing the family Nyctaginaceae and relatives, which, in the MQSST analysis also contained Limeaceae. Both the ML and MQSST recovered a strongly supported clade that consisted of the families Kewaceae, Aizoaceae, Gisekiaceae, Sarcobataceae, Agdestidaceae, Phytolaccaceae, Petiveriaceae, and Nytaginaceae (Fig. 3). Kewaceae was sister to all others, with Aizoaceae diverging first amongst the remaining members, followed by the monotypic Gisekiaceae. The next lineage to diverge is a clade containing the family Phytolaccaceae as sister to a strongly supported clade including the families Sarcobataceae and Agdestidaceae. Next, there is a strongly supported clade of Petiveriaceae+Nyctaginaceae.

The Portullugo clade containing Molluginaceae and Portulacineae was strongly supported as monophyletic. The family Molluginaceae was sister to the rest of the Portulacineae, with the divergences of Montiaceae, Basellaceae, Didieraceae and Talinaceae resolved as a grade (Fig. 3). As sister to the family Talinaceae, a clade was recovered in which the family Cactaceae was sister to a clade of Anacampserotaceae+Portulacaceae. The monophyly and placements of all families were strongly supported.

<h2>Phylogenetic resolution among and within major Caryophyllales families from IND analyses

<h3>Noncore Caryophyllales

The sampling of the noncore Caryophyllales consisted of 60 species, with at least one representative from eight of the nine families (Fig. 4). All species relationships from the IND analysis were congruent with those of the ALL MQSST with the exception of the placement of *Eriogonum longifolium*. The family Polygonaceae had the highest density of sampling with 37 taxa. The IND analysis of the ML and MQSST analyses had a final matrix occupancy of ~81% (Appendix S3). The MQSST and the ML supermatrix analyses were largely congruent, aside from the genus *Eriogonum,* where all nodes contained at least 50% gene-tree discordance, and a few relationships had low LPP support. The families of carnivorous taxa (including Ancistrocladaceae, which has reverted to be noncarnivorous) formed a clade. The four families of the NCNC were also monophyletic, but with medium LPP support and (>50%) gene-tree conflict.

<h3>Amaranthaceae/Chenopodiaceae

The results of the IND MQSST and the ALL MQSST analyses were concordant except in the position of the genus *Beta* and the species *Tidestromia lanuginosa*. Sampling consisted of 60 Amaranthaceae, 39 of the taxa were members of the former ‘Chenopodiaceae’ (Hernández-Ledesma et al., 2015; APG IV, 2016), and five Caryophyllaceae samples were used as outgroups. The choice of orthologs used in the analysis contained at least 60 taxa, resulting in 455 orthologs with approximately 15.5% missing data in the supermatrix (Appendix S3). The MQSST and ML analysis contained three discrepancies (Fig. 5), all of which were marked by a minimum of 75% gene tree discordance and imperfect LPP support at the contentious node. The ML/MQSST conflict surrounded the relationships of the genus *Beta* where it is either sister to all other Chenopodiaceae or found nested within Chenopodiaceae. Another conflict was the relationship of *Krascheninnikovia lanata* and *Suckleya suckleyana*, where the two taxa appeared as sister in the supermatrix analysis, but showed *S. suckleyana* and *K. lanata* formed a grade in the MQSST analysis (Fig. 5).

The majority of missing sequence data for the analysis was found in the clade that consists of the genus *Suaeda*, and the position of *Bienertia* as sister to *Suaeda* had a dominant alternative topology that consisted of roughly the same number of gene trees as the rest of the conflict (Fig. 5). Most of the conflict in the relationships was located at deeper nodes along the phylogeny. We found 376 of the 455 gene trees conflicted with the species tree surrounding the paraphyly of Amaranthaceae, with *Nitrophila occidentalis* and *Polynemum majus* forming a clade sister to the species that were formerly recognized as Chenopodiaceae. Although gene-tree concordance was low (~17%), there was no dominant alternative topology found among the conflicting topologies.

<h3>Cactaceae

The inferred topology from the ALL MQSST analysis was congruent with the IND MQSST analysis of Cactaceae, except for the relationship between the genera *Leuenbergeria* and *Pereskia* and the relationship of the genera *Gymnocalycium* and *Stetsonia* (Figs. 3, 6). The

Cactaceae sampling included 29 ingroup taxa and inference of the IND species tree was done using 1502 orthologs with ~19% missing data in the final supermatrix (Appendix S3). The MQSST and ML supermatrix species trees contained a high level of gene-tree conflict among many relationships (Fig. 6), including whether the nonsucculent taxa, previously circumscribed as *Pereskia* (now *Leuenbergeria* and *Pereskia*), were monophyletic or paraphyletic. High gene-tree conflict (>75%) was prevalent across many relationships including the position of *Lophophora williamsii*, the relationship of *Salmiopuntia salmiana* and *Tunilla corrugata*, and the relationship of the genus *Pereskia* with respect to the genus *Leuenbergeria*. Most of the missing data for the analysis was from the two species in the genus *Pereskia*.

<h3>Caryophyllaceae

The topologies of the all-species MQSST analysis and the IND MQSST analysis were completely concordant. The sampling across Caryophyllaceae consisted of 39 ingroup taxa and inference of the IND species tree was done using 999 orthologs, with ~17% missing data in the final supermatrix (Appendix S3). The MQSST and ML supermatrix species tree analyses resulted in congruent topologies, with perfect LPP support at almost all nodes (Fig. 7). Most genera were recovered as monophyletic, with the exception of *Arenaria,* where almost all gene trees placed *Arenaria procera* sister to *Eremogone hookeri.*

<h3>Nyctaginaceae

The ALL MQSST analysis was concordant with the IND analysis aside from the position of *Boerhavia ciliata.* The sampling across the Nyctaginaceae consisted of 37 Nyctaginaceae with four Petiveriaceae used as outgroups. The species tree was inferred using 389 orthologs with ~14% missing data for the final ML supermatrix. The MQSST and supermatrix IND analyses were largely congruent aside from the relationships among species in the genus *Boerhavia* (Fig. 8). Within *Boerhavia*, there were 111 gene trees supporting *Boerhavia coccinea* sister to *Boerhavia torreyana* and 107 gene trees supporting an alternative of *B. coccinea* sister to *Boerhavia purpurascens*. The incongruent node contains a large amount of conflicting gene-tree signal with a dominant alternative topology matching the MQSST analysis. The node supporting the monophyletic herbaceous xerophytic clade contains almost no gene-tree conflict.

<h1>DISCUSSION

<h2>Utilizing broad and dense transcriptome sampling for inference in Caryophyllales

Previous phylogenetic analyses have vastly improved our understanding of the backbone relationships of Caryophyllales (Rodman et al., 1984; Cuenod et al., 2002; Brockington et al., 2009; Schäferhoff et al., 2009; Yang et al., 2015), but strong resolution of early-diverging lineages has proven a formidable task. Here, with increased taxon sampling and larger datasets, we reconstructed most relationships with high support (Fig. 3). The higher support was found for deeper-level relationships that previously had weak or moderate support (e.g., Sarcobataceae and Agdestidaceae), as well as for new hypotheses (e.g., Stegnospermataceae as sister to Macarthuriaceae). Reassuringly, and similar to the results of other phylogenomic studies (Cannon et al., 2015; Yang et al., 2015; Pease et al., 2016), we find most relationships in the tree are concordant with previous single- or multi-gene studies. This consistency indicates that, in many cases, data sets of one or a few genes are sufficiently powerful for inferring most species relationships. While this improved resolution highlights the power of large nuclear data sets for phylogenetic inference, it is important to note that such data sets are not a phylogenetic panacea. For example, our analyses conflicted with the previously inferred monophyly of the families within the noncore Caryophyllales. Both MQSST and ML analyses found noncore Caryophyllales to be nonmonophyletic (which was weakly supported as monophyletic in Yang et al. [2015]), the MQSST placed the carnivorous Caryophyllales with zero branch length and no support as sister to the core Caryophyllales. The ML analyses weakly supported the noncarnivorous noncore as sister to the core Caryophyllales (Appendices S4–S6).

The inability of >1000 orthologs to provide statistical support for this relationship demonstrates a limitation of the current methods for phylogenetic inference with large data sets. The lack of resolution may be due to methodological limitations (e.g., model misspecification or oversimplification) or biological reality (e.g., biological processes occurred that obfuscate this relationship and leave little to no informative signal). Many relationships are the result of complex evolutionary histories that are manifested in conflict among gene-tree topologies. Although conflict makes it difficult to infer species relationships, phylotranscriptomics provides a cost-efficient means of identifying conflicting gene trees and hence potentially exposing the underlying evolutionary processes, including ILS, hybridization, and gene duplication, that are often masked when using a small number of genes. Some of these recalcitrant phylogenetic relationships may be resolved by more sophisticated methods (Olave et al., 2015), but some may never be resolved due to the complex nature of evolution and speciation (e.g., hybridization, ILS, and gene duplication and loss). The complex nature of evolution can even lead to cases of “hard polytomies” originating when lineages radiate almost simultaneously from a common ancestor.

The analyses presented here add to a growing number of phylogenomic analyses that have exposed extensive underlying gene-tree conflict (Smith et al., 2015; Pease et al., 2016; Walker et al., 2017). Methods for analyzing and incorporating this conflict are rapidly emerging (Ané et al., 2007; Leigh et al., 2008; Salichos et al., 2014; Smith et al., 2015; Kobert et al., 2016; Arcila et al., 2017). We found, as with previous studies, that gene-tree conflict was unevenly distributed. For example, clades that may have undergone a rapid radiation (e.g., Cactaceae) (Arakaki et al., 2011) exhibit more gene-tree conflict than others (e.g., Caryophyllaceae). In some cases, we found nodes with as few as 50 of 455 gene trees (~17%) supporting the ML and MQSST relationship (e.g., the position of the subfamily Polycnemoideae within the Amaranthaceae/Chenopodiaceae). However, in this case, the relationship with the next most gene-tree support, 29 of 455 (~6%), recovered the Polycnemoideae as sister to both Chenopodiaceae/Amaranthaceae.

Several instances of gene-tree conflict may have important taxonomic implications—for example, the most commonly inferred relationship from our molecular data indicate Polycnemoideae are more closely related to Chenopodiaceae, while they are morphologically more similar to Amaranthaceae and group with Amaranthaceae s.s. in molecular studies based on chloroplast gene regions (Masson and Kadereit, 2013 and references therein). Many traits in Polycnemoideae appear plesiomorphic and may have resulted from hybridization or ancestral polymorphism. Regardless of the underlying reasons, identifying relationships with high gene-tree conflict illustrates the power of large data sets to document evolutionary processes that cannot be elucidated with phylogenies containing only a few genes. Development of new methodologies for identifying and analyzing gene-tree conflict is an essential step forward for understanding species relationships.

Evaluating the patterns and causes of gene tree conflict results in a more informed and nuanced understanding of evolutionary history. For example, the earliest branches within the former genus *Pereskia* s.l. (now split into *Pereskia* and *Leuenbergeria*) displayed high levels of gene-tree conflict. Both *Pereskia* and *Leuenbergeria* share many defining morphological features, however, the species-tree inference resolved them to form a grade as previously demonstrated by Edwards et al. (2005). As genera and families are comprehensively examined in molecular studies, we may begin to better understand why some morphological features fail to match molecular phylogenies. In a broader sense, using phylogenomic data sets to understand the complex processes that may hide beneath perfect bootstrap support will add greater depth to the field of systematics, elucidating the complexities of the evolutionary processes responsible for adaptations that have shaped the world around us.

<h2>Taxonomic results for Caryophyllales

<h3>Agdestidaceae

Our analyses strongly support the sister relationship of *Sarcobatus* and *Agdesits* that has been suggested in several previous analyses, these typically with weak to moderate support (Brockington et al., 2011; Cuénoud et al., 2002; Schäferhoff et al., 2009). Given this relationship, and given the significant differences in floral morphology, habit, wood anatomy, etc. that characterize these genera, we suggest that both be treated as monogeneric families, as advocated by Hernández-Ledesma et al. (2015).

<h3>Amaranthaceae s. l.

The monophyly of the traditional Chenopodiaceae in our analyses, including its sister relationship to subfamily Polycnemoideae builds upon a growing body of evidence that suggests the broad circumscription of Amaranthaceae sensu APG IV (2016) may need to be reevaluated. Polycnemoideae is disjunctly distributed in Eurasia, America, and Australia and consists of only 13 (mostly rare) species in four genera. Considering the Eocene stem age Polycnemoideae appears as a relictual lineage (Masson and Kadereit, 2013). Molecular phylogenetic studies based on chloroplast markers and extensive sampling (Kadereit et al., 2003, 2012) as well as morphological similarities (petaloid tepals, filament tubes, 2-locular anthers; compare with table 5 of Kadereit et al. [2003]) place them closer to the Amaranthaceae s.s., while in terms of habitat preferences they are more like many members of the Chenopodiaceae. Our analysis contradicts the placement of Polycnemoideae in Amaranthaceae s.s. as proposed by Masson and Kadereit (2013) and provides evidence that it forms a clade sister to Chenopodiaceae in 17% of the gene trees. Nevertheless, some key early-diverging lineages in the Amaranthaceae s.l. clade are missing from our analyses (e.g., *Bosea* and *Charpentiera*); hence additional taxon sampling will be necessary to address these contradictory results.

<h2>Future directions for phylogenomic analyses of the Caryophyllales

Although we found strong resolution for many relationships among the Caryophyllales, our analysis highlights several key nodes with weak support that would benefit from more focused analyses. These include additional sampling of the missing Caryophyllales families as well as expanded sampling within major subclades of the order. For example, our results highlight the need for future investigation into the noncore Caryophyllales to explore the conflict at deep nodes in this area of the tree. The group has previously been recognized or treated as monophyletic (Brockington et al., 2009, 2011; Walker et al., 2017; Yang et al., 2018), and the poor resolution in our analyses hampers our understanding of key evolutionary events in this group (e.g., evolution of endosperm, production of secondary compounds, evolution of plant carnivory).

More extensive sampling within several families may also be necessary to resolve relationships and explore gene-tree conflict in several other areas of Caryophyllales phylogeny. For example, the discrepancy between the MQSST and the ML analyses in the placement of the family Limeaceae (Fig. 3; Appendices S4–S6) may be affected by the inclusion of only one species of *Limeum*. However, a phylogenetic study with greater taxon sampling of *Limeum* (Christin et al., 2011) agreed with the MQSST topology presented here. In any case, it is important to resolve the position of Limeaceae given its importance to the understanding of the complex pigmentation patterns seen in core Caryophyllales (Brockington et al., 2015; Lopez-Nieves et al., 2018). Further studies of Molluginaceae would also be valuable for their insight into C_4 evolution, as would more targeted studies of its sister clade the Portulacineae. More specific analyses using transcriptome data have helped uncover adaptive gene family expansions in Portulacineae (Wang et al., unpublished manuscript), multiple paleopolyploidy events in the carnivorous Caryophyllales (Walker et al., 2017) and are warranted to explore the convergent evolution of the many other extreme adaptations across Caryophyllales. Some of these include the evolution of cold tolerance across Caryophyllaceae and Polygonaceae, multiple origins of C_4 photosynthesis in Amaranthaceae s.l., and the evolution of drought tolerance in Nyctaginaceae, Polygonaceae, Aizoaceae, and Portulacineae.

<h2>Future directions for large-scale phylogenomic studies

Transcriptomics has emerged as a powerful tool for phylogenomics. The ever-decreasing costs of sequencing combined with improved methods for collecting plant material (Yang et al., 2017) and downstream data analysis (Dunn et al., 2013; Kocot et al., 2013; Yang and Smith, 2014; Emms and Kelly, 2015; Washburn et al., 2017) have made this a cost-efficient means for investigating systematic and evolutionary questions. To date, phylotranscriptomic analyses have been used at multiple phylogenetic levels, from genera (Pease et al., 2016; Yu et al., 2017) and large clades (Yang et al., 2015, 2018; McKain et al., 2016), to across all land plants (Wickett et al., 2014). As the size of these analyses continues to expand, so does their computational burden—a problem of critical importance for future research. This has never been more relevant for the botanical world than it is now, with the anticipated sequencing of 10,000 plant genomes (Cheng et al., 2018).

One challenge to increasing the size of phylogenomic data sets is the burden of homology identification. Here we explored a new approach that attempts to divide and conquer the daunting task of homology identification, breaking with the typical all-by-all BLAST procedure. By dividing the transcriptomes into smaller homology problems before combining homolog groups with a post-order (tip-to-root) method, based upon a previously hypothesized phylogeny, we dramatically reduced one major computational burden (i.e., the scaling an all-by-all BLAST). Additionally, accurate orthology detection is a key component of phylogenomic analyses, as demonstrated by a recent study demonstrating that two misidentified orthologs altered the species-tree topology in a >200-gene data set (Brown and Thomson, 2017). And so, this procedure also incorporated phylogenetic estimation into orthology detection (Gabaldón, 2008; Yang and Smith, 2014; Yang et al., 2015) as BLAST is not a phylogenetically informed means of inferring relationships (Smith and Pease, 2016).

This hierarchal method of homology identification relies on some previously identified phylogenetic relationships. After individual clades are clustered, each set of clusters is then combined (moving from tips to root) in an order defined by a simplified phylogeny. There are some benefits to this approach as, for example, it factors in clade-specific evolutionary history (e.g., a shift in molecular rate introduced from transition from a woody to an herbaceous life history). However, it may also introduce some bias as (1) it relies on some simplified phylogenetic relationships deep in the tree and (2) it assumes that the clustered groups form clades. If the groups clustered toward tips do not form a clade, clusters may be artificially broken up due to increased molecular distance of the included samples (i.e., distant species compared). While clustering in this manner may result in fewer homologs, this scenario is not likely to alter an inferred species topology. For example, in our analyses, the phylogenies recovered Polycnemoideae as sister to Chenopodiaceae. However, during homology inference, Polycnemoideae was a priori clustered with Amaranthaceae. The clustering did not force Polycnemoideae to be sister to Amaranthaceae, but clustering Polycnemoideae with Chenopodiaceae first may have resulted in more recovered homologs. However, as with any method, further investigation is warranted.

We found the tip-to-root method to be a powerful means of reducing the computational time spent conducting an all-by-all BLAST across the entire data set. However, clustering analyses that involve deep splits in the angiosperm tree of life tend to result in reduced data set size in terms of number of useful orthologs and homologs. For example, a comparison of the number of identified orthologs between the tip-to-root clustering in the current study and an all-by-all BLAST of the noncore Caryophyllales study of Walker et al., 2017, (that included 10 ingroup and two outgroup taxa) showed a greater number of inferred orthologs from the latter data set. Walker et al. (2017) recovered 1637 orthologs high matrix occupancy (i.e., most or all orthologs present for all taxa), whereas in the current study, 514 orthologs were recovered with high matrix occupancy. This discrepancy may be due to homologs being filtered by one of several cutoffs or systematic error due the difficulty of inferring larger homolog phylogenies.

The challenge of increasing taxa for phylogenomic scale data presents an interesting dichotomy. Increasing taxon sampling and phylogenetic breath and depth can improve accuracy and alter the inferred relationships and support. However, increased taxon sampling greatly increases the complexity and burden on each step in the inference process. Further explorations and methods will be required to fully realize the potential of these data sets and allow for their continued growth.

Homology detection and gene-tree conflict are not the only analytical burdens that future phylogenomic studies should seek to improve. Additional computational complexities such as evolutionary rate heterogeneity, distinguishing between ILS and hybridization, and improved understanding of gene duplication and loss will be important considerations for improving future phylogenomic analysis. While these are beyond the scope of this paper, the continued growth of phylogenomics portends an exciting time of evolutionary discovery.

<h1>ACKNOWLEDGEMENTS

We thank G. Stull and M.J. Grangé-Guermente for helpful comments on preparation of the manuscript; A. Summers, N. Douglas, A. Solís Magallanes, A. Clifford, S. Lavergne, T. Stoughton, N. Jensen, W. Judd, D. Warmington, S. Pedersen, K. Thiele, and J. Brittnacher for assisting with obtaining plant materials; the Cambridge University Botanic Gardens, Bureau of Land Management, U.S. Forest Service, the Navajo Nation, California State Parks, Missouri Botanical Garden, Rancho Santa Ana Botanic Garden, Desert Botanical Garden, the Kampong of the National Tropical Botanical Garden, Sukkulenten-Sammlung Zürich, Cairns Botanic Gardens, Botanischen Gartens–Technische Universität Dresden, Millennium Seed Bank, and Booderee National Park for granting us access to their plant materials; M.R.M. Marchán-Rivadeneira, L. Cortés Ortiz, V. S. Mandala, R. Mostow, M. Croley, L. Leatherman, and M. Parks for lab help, and M. Palmer at Matthaei Botanical Gardens for help with growing plants. We also thank the Associate Editor and two anonymous reviewers for comments that helped improve the manuscript. The molecular work of this study was conducted in part in the Genomic Diversity Laboratory at the University of Michigan.

<h1>FUNDING INFORMATION

Support came from University of Michigan, Oberlin College, the National Geographic Society, a Rackham Pre-doctoral Fellowship, a Nancy W. Walls award, and the National Science Foundation (DEB 1054539, DEB 1352907 and DEB 1354048).

<h1>LITERATURE CITED

Albert, V.A., S.E. Williams, and M.W. Chase. 1992. Carnivorous plants: phylogeny and structural evolution. *Science* 257: 1491–1495.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25: 3389–3402. http://nar.oxfordjournals.org/content/25/17/3389.short.

Ané, C., B. Larget, D.A. Baum, S.D. Smith, and A. Rokas. 2007. Bayesian estimation of concordance among gene trees. *Molecular Biology and Evolution* 24: 412–426.

APG IV [Angiosperm Phylogeny Group IV]. 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society* 181: 1–20.

Arakaki, M., P.-A. Christin, R. Nyffeler, A. Lendel, U. Eggli, R.M. Ogburn, E. Spriggs, et al. 2011. Contemporaneous and recent radiations of the world's major succulent plant lineages. *Proceedings of the National Academy of Sciences, USA* 108: 8379–8384. http://www.pnas.org/cgi/doi/10.1073/pnas.1100628108.

Arcila, D., G. Ortí, R. Vari, J.W. Armbruster, M.L.J. Stiassny, K.D. Ko, M.H. Sabaj, et al. 2017. Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. *Nature Ecology & Evolution* 1: 20. http://www.nature.com/articles/s41559-016-0020.

Brockington, S.F., R. Alexandre, J. Ramdial, M.J. Moore, S. Crawley, A. Dhingra, K. Hilu, et al. 2009. Phylogeny of the Caryophyllales sensu lato: revisiting hypotheses on pollination biology and perianth differentiation in the core Caryophyllales. *International Journal of Plant Sciences* 170: 627–643.

Brockington, S.F., R.H. Walker, B.J. Glover, P.S. Soltis, and D.E. Soltis. 2011. Complex pigment evolution in the Caryophyllales. *New Phytologist* 190: 854–864.

Brockington, S.F., Y. Yang, F. Gandia-Herrero, S. Covshoff, J.M. Hibberd, R.F. Sage, G.K.S. Wong, et al. 2015. Lineage-specific gene radiations underlie the evolution of novel betalain pigmentation in Caryophyllales. *New Phytologist* 207: 1170–1180.

Brown, J.M., and R.C. Thomson. 2017. Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. *Systematic Biology* 66: 517–530.

Brown, J.W., J.F. Walker, and S.A. Smith. 2017. Phyx: phylogenetic tools for unix. *Bioinformatics* 33: 1886–1888. https://academic.oup.com/bioinformatics/article/2975328/Phyx:

Cannon, S.B., M.R. McKain, A. Harkess, M.N. Nelson, S. Dash, M.K. Deyholos, Y. Peng, et al. 2015. Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes. *Molecular Biology and Evolution* 32: 193–210.

Cavieres, L.A., P. Sáez, C. Sanhueza, A. Sierra-Almeida, C. Rabert, L.J. Corcuera, M. Alberdi, and L.A. Bravo. 2016. Ecophysiological traits of Antarctic vascular plants: their importance in the responses to climate change. *Plant Ecology* 217: 343–358.

Cheng, S., M. Melkonian, S. A. Smith, S. Brockington, J. M. Archibald, P.-M. Delaux, F.-W. Li et al. 2018. 10KP: a phylodiverse genome sequencing plan. *GigaScience* 7: 1–9. doi:10.1093/gigascience/giy013

Christenhusz, M.J.M., S.F. Brockington, P.A. Christin, and R.F. Sage. 2014. On the disintegration of Molluginaceae: A new genus and family (*Kewa*, Kewaceae) segregated from *Hypertelis*, and placement of *Macarthuria* in Macarthuriaceae. *Phytotaxa* 181: 238–242.

Christin, P.A., T.L. Sage, E.J. Edwards, R.M. Ogburn, R. Khoshravesh, and R.F. Sage. 2011. Complex evolutionary transitions and the significance of C_3–C_4 intermediate forms of photosynthesis in Molluginaceae. *Evolution* 65: 643–660.

Cuenod, P., V. Savolainen, L.W. Chatrou, M. Powell, R.J. Grayer, and M.W. Chase. 2002. Molecular phylogenetics of Caryophyllales based on nuclear 18S rDNA and plastid *rbcL*, *atpB* and *matK* DNA sequences. *American Journal of Botany* 89: 132–144.

Dunn, C.W., A. Hejnol, D.Q. Matus, K. Pang, W.E. Browne, S. a Smith, E. Seaver, et al. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* 452: 745–9.

Dunn, C.W., M. Howison, F. Zapata, and G.N. Jul. 2013. Agalma: an automated phylogenomics workflow. *BMC Bioinformatics* 14: 1–17. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3840672&tool=pmcentrez&rendertype=abstract.

Eaton, D.A.R., E.L. Spriggs, B. Park, and M.J. Donoghue. 2016. Misconceptions on missing data in RAD-seq phylogenetics with a deep-scale example from flowering plants. *Systematic Biology* 0: 1–14.

Eggli, U., and R. Nyffeler. 2009. Living under temporarily arid conditions – succulence as an adaptive strategy. *Bradleya* 27: 13–36.

Emms, D.M., and S. Kelly. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* 16: 157. http://genomebiology.com/2015/16/1/157.

Fernández-Mazuecos, M., G. Mellers, B. Vigalondo, L. Sáez, P. Vargas, and B.J. Glover. 2017.

Resolving recent plant radiations: power and robustness of genotyping-by-sequencing. *Systematic Biology* 0: 1–19. http://academic.oup.com/sysbio/article/doi/10.1093/sysbio/syx062/3953673/Resolving-Recent-Plant-Radiations-Power-and.

Flowers, T.J., and T.D. Colmer. 2008. Salinity tolerance in halophytes. *New Phytologist* 179: 945–63.

Fontaine, M.C., J.B. Pease, A. Steele, R.M. Waterhouse, D.E. Neafsey, I. V. Sharakhov, S.N. Mitchell, et al. 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science* 347: 1–20.

Fu, L., B. Niu, Z. Zhu, S. Wu, and W. Li. 2012. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28: 3150–3152.

Gabaldón, T. 2008. Large-scale assignment of orthology: back to phylogenetics? *Genome Biology* 9: 235. http://genomebiology.biomedcentral.com/articles/10.1186/gb-2008-9-10-235.

Galtier, N., and V. Daubin. 2008. Dealing with incongruence in phylogenomic analyses. *Philosophical Transactions of the Royal Society of London, B, Biological sciences* 363: 4023–4029.

Givnish, T.J. 2015. New evidence on the origin of carnivorous plants. *Proceedings of the National Academy of Sciences, USA* 112: 10–11.

Hernández-Ledesma, P., W.G. Berendsohn, T. Borsch, S. Von Mering, H. Akhani, S. Arias, I. Castañeda-Noa, et al. 2015. A taxonomic backbone for the global synthesis of species diversity in the angiosperm order Caryophyllales. *Willdenowia* 45: 281–383.

Jarvis, E.D., S. Mirarab, A.J. Aberer, B. Li, P. Houde, C. Li, S.Y.W. Ho, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* 346: 1320–1331. http://www.sciencemag.org/cgi/doi/10.1126/science.1253451.

Kadereit, G., D. Ackerly, and M.D. Pirie. 2012. A broader model for C_4 photosynthesis evolution in plants inferred from the goosefoot family (Chenopodiaceae s.s.). *Proceedings of the Royal Society, B, Biological Sciences* 279: 3304–3311. http://rspb.royalsocietypublishing.org/cgi/doi/10.1098/rspb.2012.0440.

Kadereit, G., T. Borsch, K. Weising, and H. Freitag. 2003. Phylogeny of Amaranthaceae and Chenopodiaceae and the evolution of C_4 photosynthesis. *International Journal of Plant*

Sciences 164: 959–986.

Katoh, K., and D.M. Standley. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* 30: 772–80.

Kobert, K., L. Salichos, A. Rokas, and A. Stamatakis. 2016. Computing the internode certainty and related measures from partial gene trees. *Molecular Biology and Evolution* 33: 1606–1617.

Kocot, K.M., M.R. Citarella, L.L. Moroz, and K.M. Halanych. 2013. PhyloTreePruner: A phylogenetic tree-based approach for selection of orthologous sequences for phylogenomics. *Evolutionary Bioinformatics* 2013: 429–435.

Leigh, J.W., E. Susko, M. Baumgartner, and A.J. Roger. 2008. Testing congruence in phylogenomic analysis. *Systematic Biology* 57: 104–115. doi:10.1080/10635150801910436.

Lopez-Nieves, S., Y. Yang, A. Timoneda, M. Wang, T. Feng, S.A. Smith, S.F. Brockington, and H.A. Maeda. 2018. Relaxation of tyrosine pathway regulation underlies the evolution of betalain pigmentation in Caryophyllales. *New Phytologist* 217: 896–908. http://doi.wiley.com/10.1111/nph.14822.

Maddison, W.P. 1997. Gene trees in species trees. *Systematic Biology* 46: 523–536.

Masson, R., and G. Kadereit. 2013. Phylogeny of Polycnemoideae (Amaranthaceae): implications for biogeography, character evolution and taxonomy. *Taxon* 62: 100–111.

McKain, M.R., H. Tang, J.R. McNeal, S. Ayyampalayam, J.I. Davis, C.W. dePamphilis, T.J. Givnish, et al. 2016. A phylogenomic assessment of ancient polyploidy and genome evolution across the Poales. *Genome Biology and Evolution* 8: 1150–1164.

Mirarab, S., and T. Warnow. 2015. ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* 31: i44–i52.

Moore, A.J., J. M. De Vos, L.P. Hancock, E. Goolsby, and E.J. Edwards. 2017. Targeted enrichment of large gene families for phylogenetic inference: phylogeny and molecular evolution of photosynthesis genes in the *Portullugo* (Caryophyllales). *Systematic Biology*: syx078. https://doi.org/10.1093/sysbio/syx078.

Moray, C., E.W. Goolsby, and L. Bromham. 2016. The phylogenetic association between salt tolerance and heavy metal hyperaccumulation in angiosperms. *Evolutionary Biology* 43: 119–130.

Olave, M., L.J. Avila, J.W. Sites Jr., and M. Morando. 2015. Model-based approach to test hard polytomies in the *Eulaemus* clade of the most diverse South American lizard genus *Liolaemus* (Liolaemini, Squamata). *Zoological Journal of the Linnean Society* 174: 169–184.

Pease, J.B., J.W. Brown, J.F. Walker, C.E. Hinchliff, and S.A. Smith. 2018. Quartet Sampling distinguishes lack of support from conflicting support in the green plant tree of life. *American Journal of Botany* 105. doi:10.1002/ajb2.1016.

Pease, J.B., D.C. Haak, M.W. Hahn, and L.C. Moyle. 2016. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. *PLOS Biology* 14: 1–24.

Price, M.N., P.S. Dehal, and A.P. Arkin. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5: 3.

Rettig, J.H., H.D. Wilson, and J.R. Manhart. 1992. Phylogeny of the Caryophyllales: gene sequence data. *Taxon* 41: 201–209.

Rodman, J.E., M.K. Oliver, R.R. Nakamura, U. James, J.E. Rodman, and A.H. Bledsoe. 1984. A taxonomic analysis and revised classification of Centrospermae. *Systematic Botany* 9: 297–323.

Rokas, A., B.L. Williams, N. King, and S.B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425: 798–804.

Sage, R.F. 2017. A portrait of the C_4 photosynthetic family on the 50th anniversary of its discovery: species number, evolutionary lineages, and Hall of Fame. *Journal of Experimental Botany* 68: e11–e28.

Sage, R.F., P.-A. Christin, and E.J. Edwards. 2011. The C_4 plant lineages of planet Earth. *Journal of Experimental Botany* 62: 3155–3169. http://jxb.oxfordjournals.org/lookup/doi/10.1093/jxb/err048.

Sajeva, M., and J.D. Mauseth. 1991. Leaf-like structure in the photosynthetic, succulent stems of cacti. *Annals of Botany* 68: 405–411.

Salichos, L., A. Stamatakis, and A. Rokas. 2014. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. *Molecular Biology and Evolution* 31: 1261–1271.

Sayyari, E., and S. Mirarab. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. *Molecular Biology and Evolution* 33: 1654–1668.

Schäferhoff, B., K.F. Müller, and T. Borsch. 2009. *Caryophyllales* phylogenetics: disentangling

Phytolaccaceae and *Molluginaceae* and description of *Microteaceae* as a new isolated family. *Willdenowia* 39: 209–228.

Seo, T.K. 2008. Calculating bootstrap probabilities of phylogeny using multilocus sequence data. *Molecular Biology and Evolution* 25: 960–971.

Shen, X.X, C.T. Hittinger, and A. Rokas. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecology & Evolution* 1: 1–10. http://dx.doi.org/10.1038/s41559-017-0126.

Smith, S. A., and J. W. Brown. 2018. Constructing a broadly inclusive seed plant phylogeny. *American Journal of Botany* 105 doi: 10.1002/ajb2.1019.

Smith, S.A., J.W. Brown, and J.F. Walker. 2017. So many genes, so little time: comments on divergence-time estimation in the genomic era. *bioRxiv:* 114975. https://doi.org/10.1101/114975

Smith, S.A., J.W. Brown, Y. Yang, R. Bruenn, C.P. Drummond, S.F. Brockington, J.F. Walker, et al. 2018. Disparity, diversity, and duplications in the Caryophyllales. *New Phytologist* 217: 836–854. http://doi.wiley.com/10.1111/nph.14772.

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., and J.B. Pease. 2016. Heterogeneous molecular processes among the causes of how sequence similarity scores can fail to recapitulate phylogeny. *Briefings in Bioinformatics* 18: bbw034. https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbw034.

Smith, S.A., N.G. Wilson, F.E. Goetz, C. Feehery, S.C.S. Andrade, G.W. Rouse, G. Giribet, and C.W. Dunn. 2011. Resolving the evolutionary relationships of molluscs with phylogenomic tools. *Nature* 480: 364–367. http://www.ncbi.nlm.nih.gov/pubmed/22031330%5Cnhttp://www.nature.com/doifinder/10.1038/nature10526.

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

Stevens, P. 2001 [onward]. Angiosperm Phylogeny Website [more or less continuously updated]. http://www.mobot.org/MOBOT/research/APweb/ [accessed 1 January 2015].

Thulin, M., A.J. Moore, H. El-Seedi, A. Larsson, P.A. Christin, and E.J. Edwards. 2016.

Phylogeny and generic delimitation in Molluginaceae, new pigment data in Caryophyllales, and the new family Corbichoniaceae. *Taxon* 65: 775–793.

Van Dongen, S. 2000. Graph clustering by flow simulation. University of Utrecht, Utrecht, Netherlands.

Walker, J.F., J.W. Brown, and S.A. Smith. 2018. Analyzing contentious relationships and outlier genes in phylogenomics. *bioRxiv*, 11574. doi:10.1101/115774.

Walker, J.F., Y. Yang, M.J. Moore, J. Mikenas, A. Timoneda, S.F. Brockington, and S.A. Smith. 2017. Widespread paleopolyploidy, gene tree conflict, and recalcitrant relationships among the carnivorous Caryophyllales. *American Journal of Botany* 104: 858–867.

Walker, J.F., M.J. Zanis, and N.C. Emery. 2014. Comparative analysis of complete chloroplast genome sequence and inversion variation in *Lasthenia burkei* (Madieae, Asteraceae). *American Journal of Botany* 101: 722–729.

Washburn, J.D., J.C. Schnable, G.C. Conant, T.P. Brutnell, Y. Shao, Y. Zhang, M. Ludwig, et al. 2017. Genome-guided phylo-transcriptomic methods and the nuclear phylogentic tree of the Paniceae grasses. *Scientific Reports* 7: 13528. http://dx.doi.org/10.1038/s41598-017-13236-z.

White, P.J., H.C. Bowen, M.R. Broadley, H.A. El-Serehy, K. Neugebauer, A. Taylor, J.A. Thompson, and G. Wright. 2017. Evolutionary origins of abnormally large shoot sodium accumulation in nonsaline environments within the Caryophyllales. *New Phytologist* 214: 284–293.

Wickett, N.J., S. Mirarab, N. Nguyen, T. Warnow, E. Carpenter, N. Matasci, S. Ayyampalayam, et al. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences, USA* 111: E4859–E4868. http://www.pnas.org/content/111/45/E4859.abstract.

Yang, Y., M.J. Moore, S.F. Brockington, J. Mikenas, J. Olivieri, J.F. Walker, and S.A. Smith. 2018. Improved transcriptome sampling pinpoints 26 ancient and more recent polyploidy events in Caryophyllales, including two allopolyploidy events. *New Phytologist* 217: 855-870. http://doi.wiley.com/10.1111/nph.14812.

Yang, Y., M.J. Moore, S.F. Brockington, D.E. Soltis, G.K.-S. Wong, E.J. Carpenter, Y. Zhang, et al. 2015. Dissecting molecular evolution in the highly diverse plant clade Caryophyllales using transcriptome sequencing. *Molecular Biology and Evolution* 32: 8.

Yang, Y., M.J. Moore, S.F. Brockington, A. Timoneda, T. Feng, H.E. Marx, J.F. Walker, and S.A. Smith. 2017. An efficient field and laboratory workflow for plant phylotranscriptomic projects. *Applications in Plant Sciences* 5: 3.

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

Yu, Y., Q. Xiang, P.S. Manos, D.E. Soltis, P.S. Soltis, B.H. Song, S. Cheng, et al. 2017. Whole-genome duplication and molecular evolution in *Cornus* L. (Cornaceae)—Insights from transcriptome sequences. *PLOS ONE* 12: 1–21.

FIGURE 1. Transcriptome sampling across the Caryophyllales. Species level tree of the Caryophyllales from Smith et al. (2018), collapsed to genus level. Branches of genera included in the study are highlighted in red, and circles at the tips are proportionate to the number of samples from a given genus.

FIGURE 2. Representation of the post-order clustering method. Diagram of the general order in which clustering was performed, based upon a synthesis of previous phylogenies (Stevens, 2001). For outgroups, the name of the genus was given and, for Caryophyllales, the family name. In brackets, the number of individuals from that family sequenced is listed. Semi-circles represent tip-level all-by-all clustering, and full circles represent node level clustering, with numbers representing number of clustering procedures performed to create the given node level cluster.

FIGURE 3. Caryophyllales phylogeny inferred from 305 transcriptomes. The maximum quartet support species tree inferred from 305 transcriptomes. Branches are colored in a gradient to represent support, with cooler colors (blue) representing strong support and warmer colors (red) representing weak support. In the bottom left corner, a tree shows the relationships among major families, with stars depicting major family level findings.

FIGURE 4. Inferred species relationships among taxa in the families of the noncore Caryophyllales. Phylogeny inferred using maximum likelihood (ML) from the concatenated data set of the 514 inferred orthologs across the noncore Caryophyllales. Branches in red represent conflict with the maximum quartet support species tree. Gene-tree conflict is represented as pie

charts on the ML tree, blue indicates proportion of gene trees concordant with the ML tree topology, green indicates the most common alternative gene-tree topology, red indicates conflicting gene trees with other alternative topologies, and grey indicates sampling was missing for the gene tree to infer a given relationship. *Photo credits: Oxytheca porfoliata*, Stan Shebs; *Fagopyrum vesculentum*, Kurt Stüber; *Frankenia laevis*, Ghislain118; and *Nepenthes alata*, *Drosophyllum lusitanicum*, and *Dionaea muscipula* (trap and flower), Joe Walker. Licenses and location of original photographs can be found in Appendix S7.

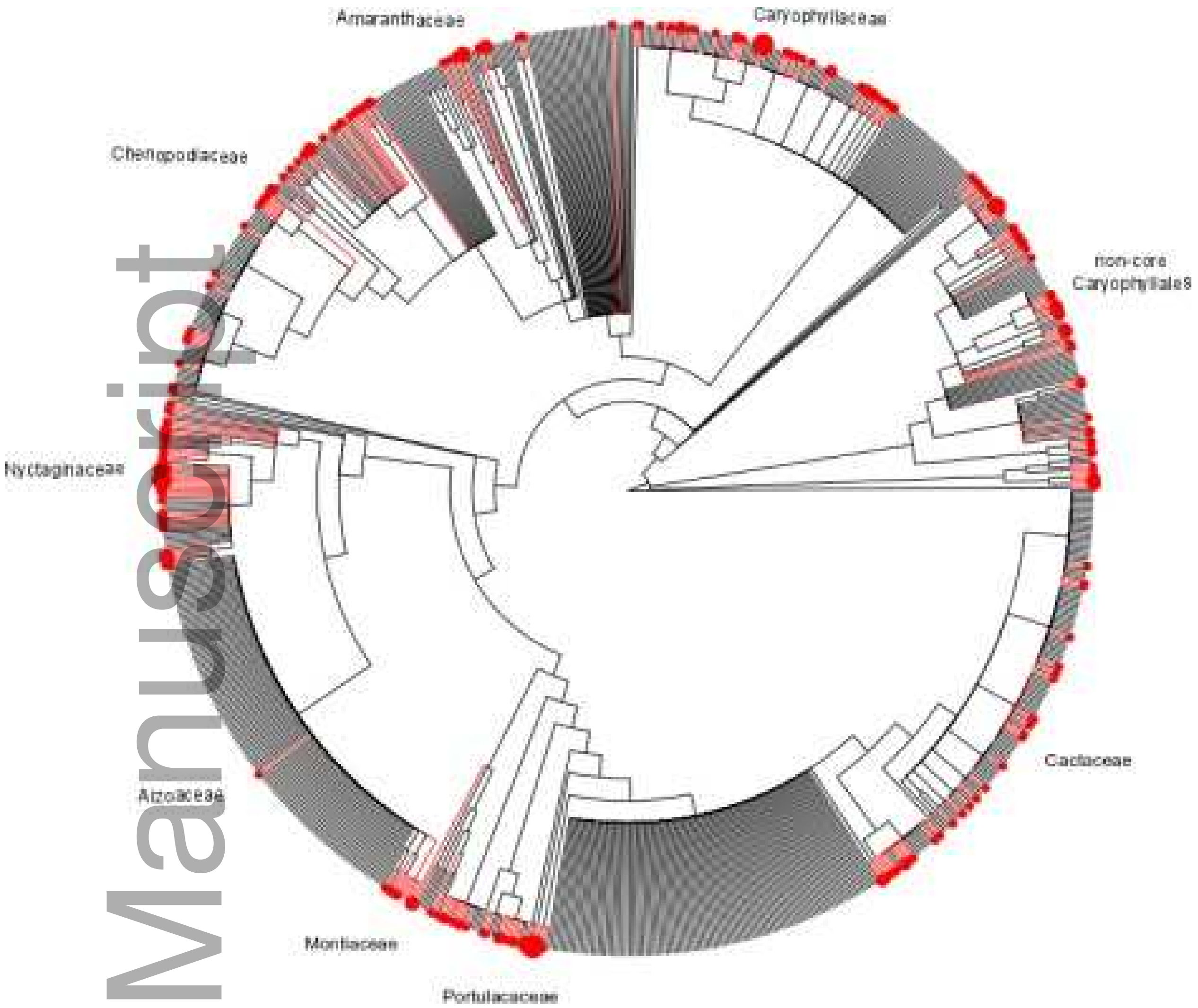
FIGURE 5. Inferred species relationships among taxa in the families Amaranthaceae and Chenopodiaceae. Phylogeny inferred using maximum likelihood (ML) from the concatenated data set of the 455 inferred orthologs across the Amaranthaceae and Chenopodiaceae. Branches in red represent conflict with the maximum quartet support species tree. Gene-tree conflict is represented as pie charts on the ML tree, blue indicates proportion of gene trees concordant with the ML tree topology, green indicates the most common alternative gene tree topology, red indicates conflicting gene trees with other alternative topologies, and grey indicates sampling was missing for the gene tree to infer a given relationship. *Photo credits: Grayia spinosa*, Stan Shebs; *Spinacia oleraceae*, Victor M. Vincent Selvas; *Beta vulgaris*, Evan Amos; *Nitrophila occidentalis*, Michael J. Moore; *Amaranthus tricolor*, Kurt Stueber. Licenses and location of original photographs can be found in Appendix S7.

FIGURE 6. Inferred species relationships among taxa in the Cactaceae. Phylogeny inferred using maximum likelihood (ML) from the concatenated data set of the 1502 inferred orthologs across the Cactaceae. Branches in red represent conflict with the maximum quartet support species tree. Gene-tree conflict is represented as pie charts on the ML tree, blue indicates proportion of gene trees concordant with the ML tree topology, green indicates the most common alternative gene tree topology, red indicates conflicting gene trees with other alternative topologies, and grey indicates sampling was missing for the gene tree to infer a given relationship. *Photo credits: Ferocactus latispinus* and *Opuntia arenaria*, Lucas C. Majure; *Pereskia grandiflora*, Kurt Stüber. Licenses and location of original photographs can be found in Appendix S7.

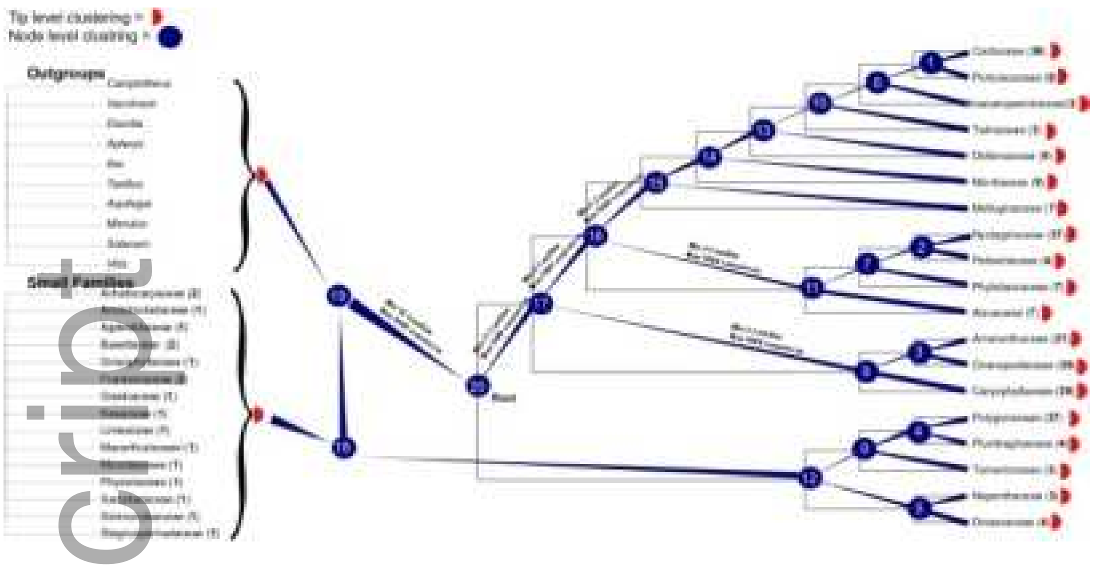
FIGURE 7. Inferred species relationships among taxa in the Caryophyllaceae. Phylogeny inferred using maximum likelihood (ML) from the concatenated data set of the 999 inferred orthologs across the Caryophyllaceae. Gene-tree conflict is represented as pie charts on the ML

tree, blue indicates proportion of gene trees concordant with the ML tree topology, green indicates the most common alternative gene tree topology, red indicates conflicting gene trees with other alternative topologies, and grey indicates sampling was missing for the gene tree to infer a given relationship. *Photo credits: Cerastium arvense*, Walter Siegmund; *Colobanthus quitensis*, Liam Quinn; *Dianthus caryophyllus*, Pagemoral; *Silene latifolia*, Walter Siegmund. Licenses and location of original photographs can be found in Appendix S7.

FIGURE 8. Inferred species relationships among taxa in the Nyctaginaceae. Phylogeny inferred using maximum likelihood (ML) from the concatenated data set of the 389 inferred orthologs across the Nyctaginaceae. Branches in red represent conflict with the maximum quartet support species tree. Gene-tree conflict is represented as pie charts on the ML tree, blue indicates proportion of gene trees concordant with the ML tree topology, green indicates the most common alternative gene tree topology, red indicates conflicting gene trees with other alternative topologies, and grey indicates sampling was missing for the gene tree to infer a given relationship. *Photo credits: Nyctaginia capitata*, *Mirabilis multiflora*, and *Abronia umbellata*, Michael J. Moore; *Pisonia umbellifera*, Forest and Kim Starr. Licenses and location of original photographs can be found in Appendix S7.



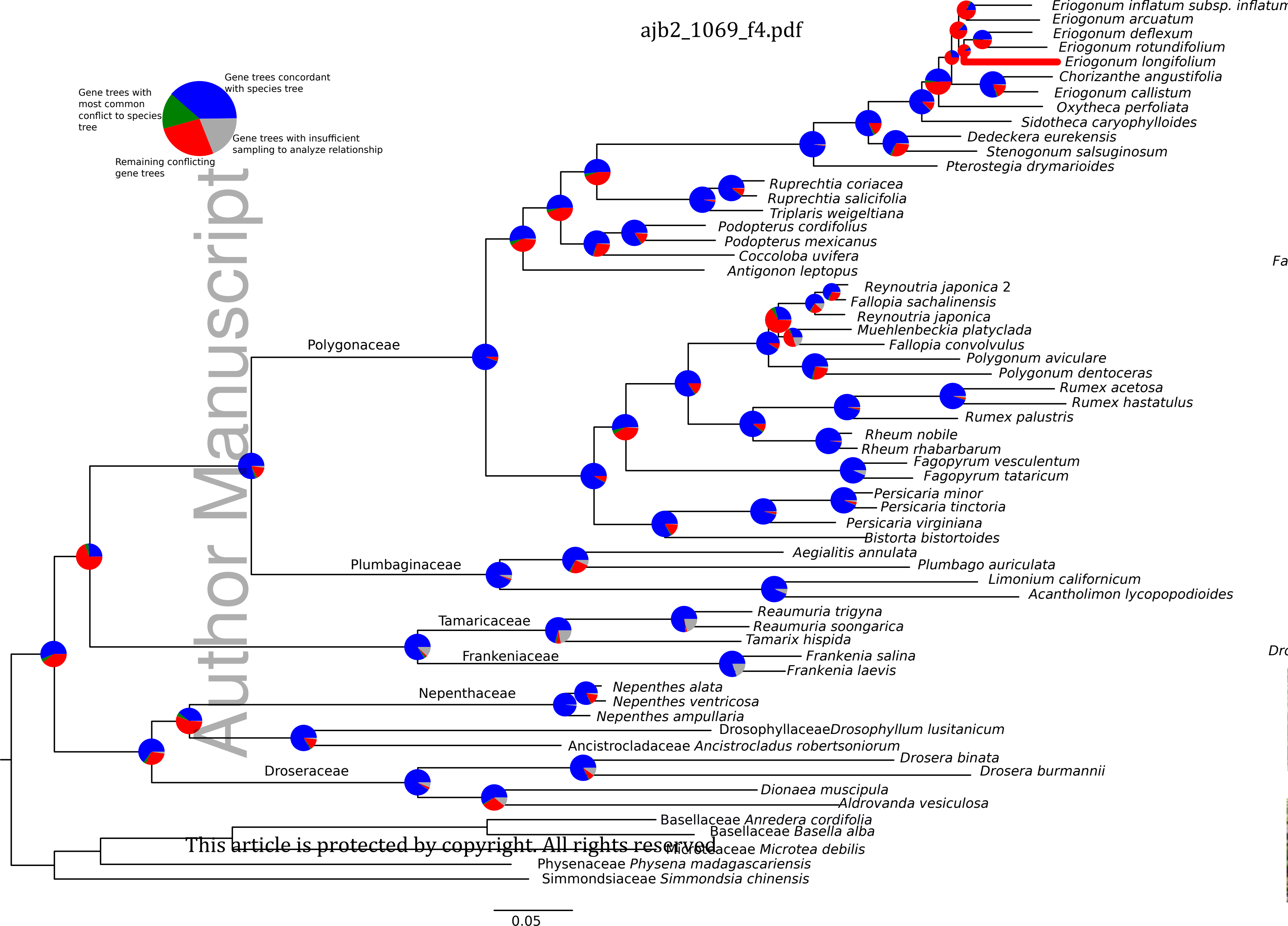
ajb2_1069_f1.png



ajb2_1069_f2.png

1.00
0.00
Matucana aurantiaca
Echinopsis aurea
Stetsonia coryne
Gymnocalycium mihanovichii
Eriosyce wageneckii
Pachycereus gatesii
Stenocereus yunckeri
Echinocereus pectinatus
Hylocereus lemairei
Peniocereus cuixmalensis
Rhipsalis baccifera subsp baccifera
Lophophora williamsii
Coryphantha maiztablensis
Ariocarpus retusus
Ferocactus latispinus
Astrophytum asterias
Astrophytum myriostigma
Copiapoa desertorum
Maihuenia poeppigii
Opuntia streptacantha
Opuntia bravoana
Opuntia cochenillifera
Opuntia arenaria
Salmiopuntia salmiana
Tunilla corrugata
Tacinga lilae
Maihueniopsis conoidea
Pterocactus tuberosus
Tephrocactus bonnieae
Tephrocactus articulatus
Grusonia bradtiana
Pereskia grandifolia
Pereskia aculeata
Leuenbergeria guamacho
Leuenbergeria lychnidiflora
Leuenbergeria bleo
Cactaceae
Portulaca pilosa
Portulaca suffrutescens
Portulaca amilis
Portulaca grandiflora
Portulaca umbraticola
Portulaca cryptopetala
Portulaca molokiniensis
Portulaca oleracea
Portulacaceae
Grahamia kurtzii
Anacampseros filamentosa
Talinopsis frutescens
Anacampserotaceae
Talinum portulacifolium
Talinum fruticosum
Talinum paniculatum
Talinaceae
Alluaudia procera
Alluaudiopsis marnieriana
Alluaudia humbertii
Alluaudia dumosa
Didierea madagascariensis
Didierea trollii
Decarya madagascariensis
Ceraria pygmaea
Portulacaria afra
Didiereaceae
Anredera cordifolia
Basella alba
Basellaceae
Claytonia virginica
Claytonia nevadensis
Montia chamissoi
Lewisia nevadensis
Calandrinia grandiflora or procumbens
Calyptridium umbellatum
Calyptridium pygmaeum
Cistanthe grandiflora
Phemeranthus parviflorus
Montiaceae
Suessenguthiella caespitosa
Pharnaceum exiguum
Mollugo cerviana
Mollugo nudicaulis
Glinus lotoides var virens
Mollugo verticillata
Mollugo pentaphylla
Molluginaceae
*
Centrospermae
Agdestidaceae
Sarcobataceae
Phytolacaceae
Petiveraceae
Nyctaginaceae
Gisekiaceae
Aizoaceae
Kewaceae
Anacampserotaceae
Portulaceae
Cactacaee
Talinaceae
Didiereaceae
Basellaceae
Montiaceae
Molluginaceae
Limeaceae
Polycmoideae
Chenopodiaceae
Amaranthaceae
Achatocarpaceae
Caryophyllaceae
Macarthuriaceae
Stegnospermataceae
Microteaceae
Physenaceae
Simmondsiaceae
Tamaricaceae
Frankeniaceae
Polygonaceae
Plumbaginaceae
Drosophyllaceae
Ancistrocladaceae
Nepenthaceae
Droseraceae
*
*
*
*
Boerhavia coccinea
Boerhavia purpurascens
Boerhavia torreyana
Boerhavia burbidgeana
Okenia hypogea
Boerhavia ciliata
Anulocaulis leiosolenus
Anulocaulis annulatus
Anulocaulis eriosolenus
Nyctaginia capitata
Cyphomeris gypsophiloides
Allionia incarnata
Commicarpus scandens
Mirabilis jalapa
Mirabilis pringlei
Mirabilis multiflora
Abronia glabrifolia
Abronia bigelovii
Abronia carletonii
Abronia fragrans
Abronia latifolia
Abronia umbellata
Abronia maritima
Tripterocalyx carneus
Tripterocalyx crux-maltae
Acleisanthes lanceolata
Acleisanthes purpusiana
Acleisanthes chenopodioides
Acleisanthes obtusa
Bougainvillea stipitata
Bougainvillea spectabilis
Neea psychotrioides
Guapira obtusata
Pisonia aculeata
Pisonia umbellifera
Colignonia ovalifolia
Salpianthus purpurascens
Nyctaginaceae
Rivina humilis
Hilleria latifolia
Trichostigma octandrum
Petiveria alliacea
Monococcus echinophorus
Seguieria aculeata
Petiveriaceae
Anisomeria littoralis
Phytolacca dioica
Phytolacca americana
Phytolacca bogotensis
Ercilla volubilis
Phytolaccaceae
Agdestis clematidea
Agdestidaceae
Sarcobatus vermiculatus
Sarcobataceae
Gisekia pharnaceoides
Gisekiaceae
Sesuvium portulacastrum
Sesuvium verrucosum
Cypselea humifusa
Zaleya pentandra
Trianthema portulacastrum
Mesembryanthemum crystallinum
Delosperma echinatum
Aizoaceae
Kewa bowkeriana
Kewaceae
Limeum aethiopicum
Limeaceae
Zuckia brandegeei
Grayia spinosa
Extriplex californica
Stutzia covillei
Proatriplex pleiantha
Atriplex prostrata
Atriplex hortensis
Atriplex rosea
Atriplex sp
Chenopodium quinoa
Chenopodiastrum murale
Chenopodium giganteum
Oxybasis rubra
Spinacia turkestanica
Spinacia tetrandra
Spinacia oleracea
Krascheninnikovia lanata
Suckleya suckleyana
Agriophyllum squarrosum
Corispermum hyssopifolium
Beta maritima
Beta vulgaris
Salicornia europaea
Salicornia pacifica
Tecticornia pergranulata
Arthrocaulon macrostachyum
Kalidium cuspidatum
Allenrolfea sp. (predicted)
Suaeda linearis
Suaeda maritima
Suaeda fruticosa
Bienertia sinuspersici
Haloxylon ammodendron
Halogeton glomeratus
Anabasis articulata
Kali collina
Caroxylon vermiculatum
Eokochia saxicola
Bassia scoparia
Chenopodiaceae
Polycnemum majus
Nitrophila occidentalis
Amaranthaceae
Alternanthera philoxeroides
Alternanthera brasiliana
Alternanthera sessilis
Alternanthera caracasana
Alternanthera tenella
Tidestromia lanuginosa
Blutaparon vermiculare
Gossypianthus lanuginosus
Guilleminea densa var aggregata
Froelichia floridana
Nelsia quadrangula
Iresine rhizomatosa
Iresine arbuscula
Aerva javanica
Aerva lanata
Amaranthus tricolor
Amaranthus cruentus
Amaranthus retroflexus
Amaranthus hypochondriacus
Amaranthaceae
Phaulothamnus spinescens
Achatocarpus gracilis
Achatocarpaceae
*
†♦
Centrospermae
†♦
Cerastium alpinum var lanatum
Cerastium arvense
Cerastium fontanum ssp vulgare
Pseudostellaria heterophylla
Lepyrodiclis stellarioides
Arenaria serpyllifolia
Schiedea globosa
Schiedea membranacea
Honckenya peploides
Scleranthus polycarpos
Colobanthus quitensis
Velezia rigida
Dianthus caryophyllus
Gypsophila repens
Saponaria officinalis
Silene noctiflora
Silene conica
Silene latifolia
Silene vulgaris
Silene acaulis ssp subacaulescens
Silene paradoxa
Agrostemma githago
Arenaria procera
Eremogone hookeri ssp desertorum
Spergularia media
Spergularia marina
Cardionema ramosissimum
Illecebrum verticillatum
Drymaria subumbellata
Drymaria cordata
Polycarpon tetraphyllum
Polycarpaea repens
Paronychia fastigiata
Paronychia drummondii
Paronychia jamesii
Herniaria glabra
Herniaria latifolia
Telephium imperati
Corrigiola litoralis
Caryophyllaceae
Stegnosperma halimifolium
Stegnospermataceae
Macarthuria australis
Macarthuriaceae
Microtea debilis
Microteaceae
Physena madagascariensis
Physenaceae
Simmondsia chinensis
Simmondsiaceae
Centrospermae
Nepenthes ventricosa
Nepenthes alata
Nepenthes ampullaria
Nepenthaceae
Ancistrocladus robertsonianum
Ancistrocladaceae
Drosophyllum lusitanicum
Drosophyllaceae
Drosera burmannii
Drosera binata
Dionaea muscipula
Aldrovanda vesiculosa
Droseraceae
Eriogonum inflatum subsp inflatum
Eriogonum arcuatum
Eriogonum longifolium or texanum
Eriogonum rotundifolium
Eriogonum deflexum
Eriogonum callistum
Chorizanthe angustifolia
Oxytheca perfoliata
Sidotheca caryophylloides
Stenogonum salsuginosum
Dedeckera eurekensis
Pterostegia drymarioides
Ruprechtia salicifolia
Ruprechtia coriacea
Triplaris weigeltiana
Podopterus cordifolius
Podopterus mexicanus
Coccoloba uvifera
Antigonon leptopus
Reynoutria japonica 2
Fallopia sachalinensis
Reynoutria japonica
Fallopia convolvulus
Muehlenbeckia platyclada
Polygonum dentoceras
Polygonum aviculare
Rumex hastatulus
Rumex acetosa
Rumex palustris
Rheum nobile
Rheum rhabarbarum
Fagopyrum tataricum
Fagopyrum esculentum
Persicaria minor
Persicaria tinctoria
Persicaria virginiana
Bistorta bistortoides
Polygonaceae
Aegialitis annulata
Plumbago auriculata
Acantholimon lycopopodioides
Limonium californicum
Plumbaginaceae
Reaumuria trigyna
Reaumuria soongarica
Tamarix hispida
Tamaricaceae
Frankenia salina
Frankenia laevis
Frankeniaceae
noncore Caryophyllales
Davidia involucrata
Camptotheca acuminata
Viccinium corrymbosum
Actinidia deliciosa
Ilex paraguariensis
Mimulus guttatus
Solanum lycopersicum
Outgroups
3.0

ajb2_1069_f4.pdf



Oxytheca perfoliata

Fagopyrum vesculentum

Frankenia laevis

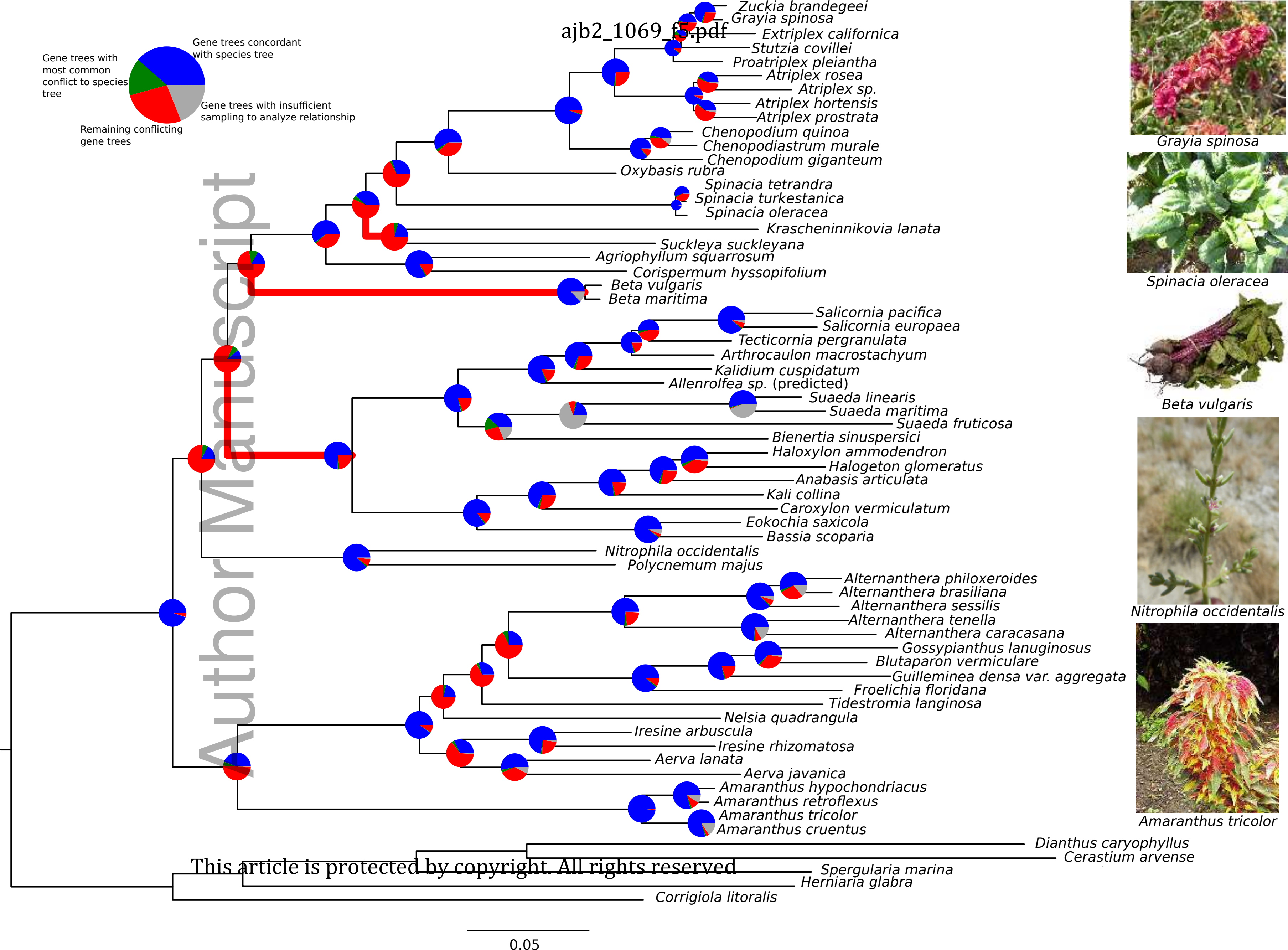
Nepenthes alata

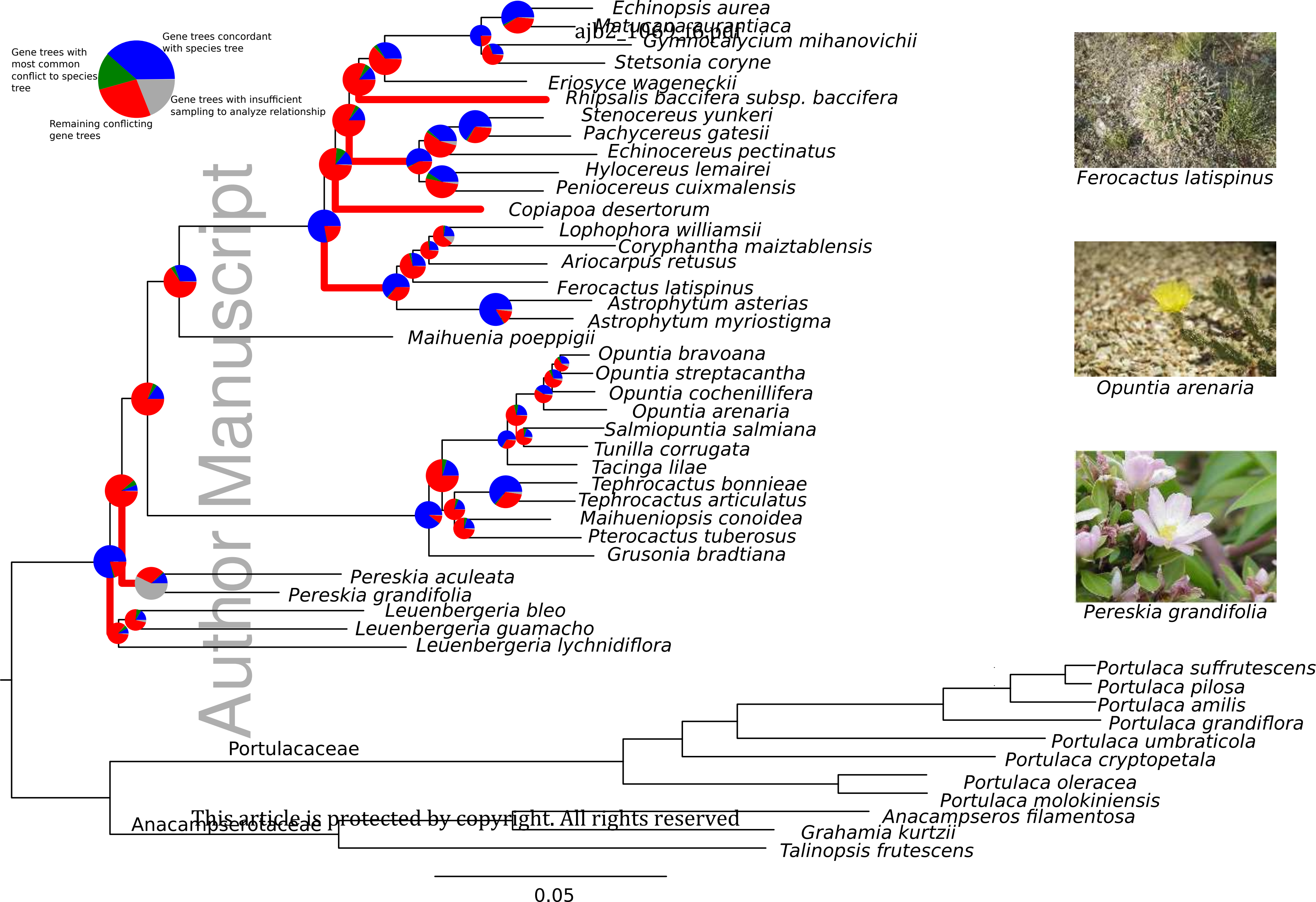
Drosophyllum lusitanicum

Dionaea muscipula

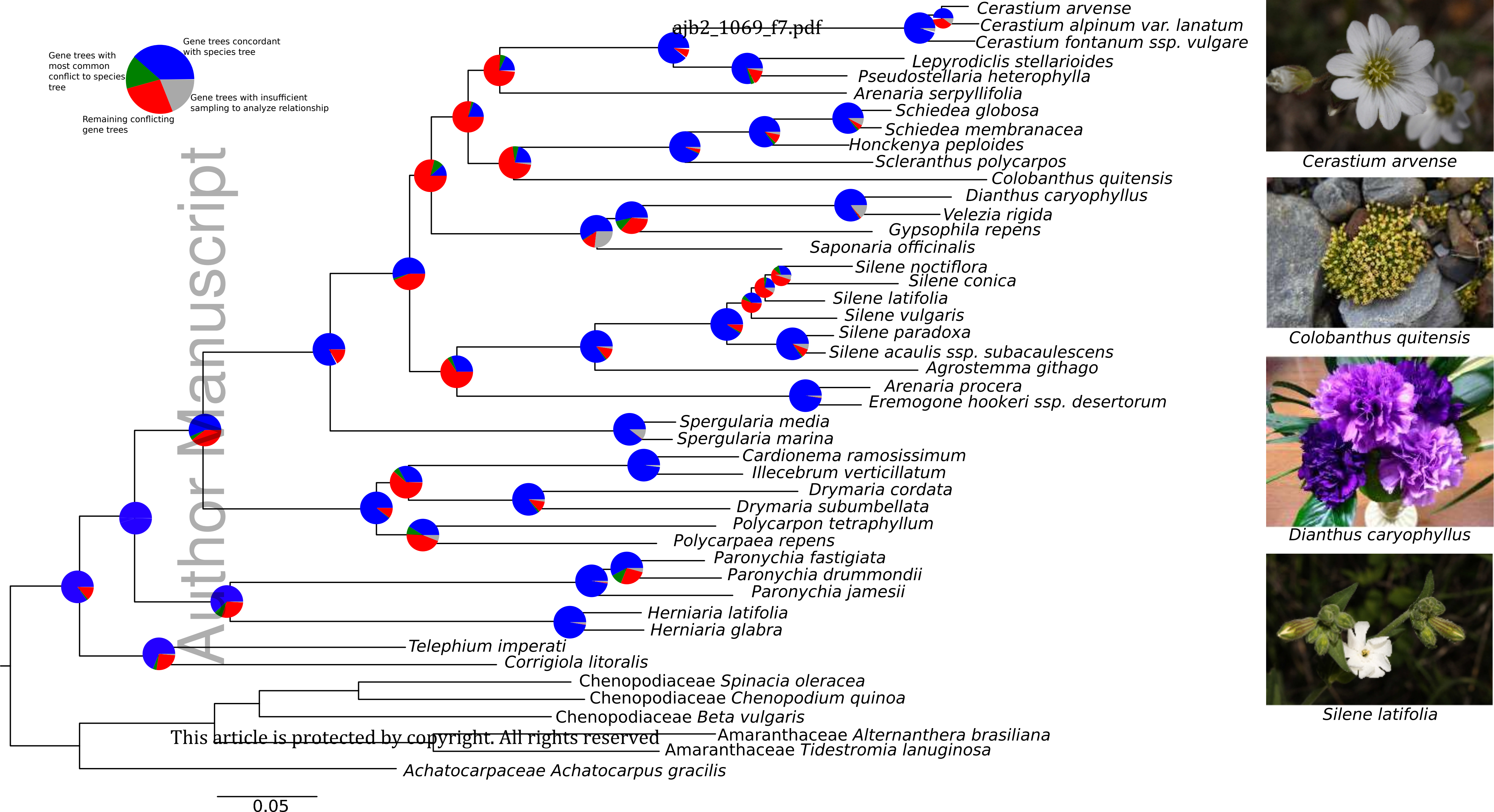
Dionaea muscipula

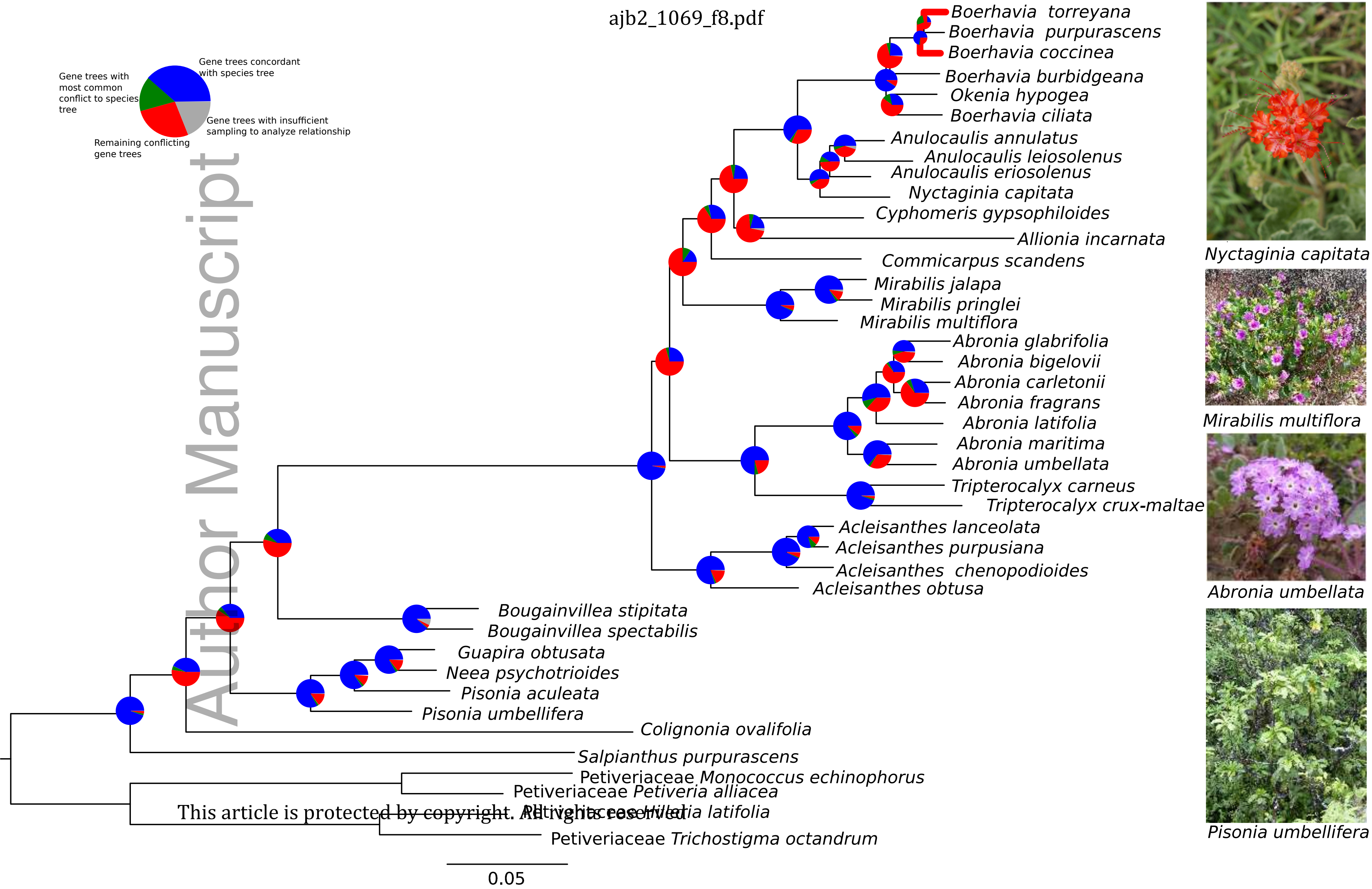
ajb2_1069_f5.pdf




Gene trees concordant with species tree
Gene trees with most common conflict to species tree
Gene trees with insufficient sampling to analyze relationship
Remaining conflicting gene trees
Echinopsis aurea
Matucana aurantiaca
Gymnocalycium mihanovichii
Stetsonia coryne
Eriosyce wageneckii
Rhipsalis baccifera subsp. baccifera
Stenocereus yunkeri
Pachycereus gatesii
Echinocereus pectinatus
Hylocereus lemairei
Peniocereus cuixmalensis
Copiapoa desertorum
Lophophora williamsii
Coryphantha maiztablensis
Ariocarpus retusus
Ferocactus latispinus
Astrophytum asterias
Astrophytum myriostigma
Maihuenia poeppigii
Opuntia bravoana
Opuntia streptacantha
Opuntia cochenillifera
Opuntia arenaria
Salmiopuntia salmiana
Tunilla corrugata
Tacinga lilae
Tephrocactus bonnieae
Tephrocactus articulatus
Maihueniopsis conoidea
Pterocactus tuberosus
Grusonia bradtiana
Pereskia aculeata
Pereskia grandifolia
Leuenbergeria bleo
Leuenbergeria guamacho
Leuenbergeria lychnidiflora
Portulaca suffrutescens
Portulaca pilosa
Portulaca amilis
Portulaca grandiflora
Portulaca umbraticola
Portulaca cryptopetala
Portulacaceae
Portulaca oleracea
Portulaca molokiniensis
Anacampseros filamentosa
Anacampserotaceae
Grahamia kurtzii
Talinopsis frutescens
0.05
Ferocactus latispinus
Opuntia arenaria
Pereskia grandifolia

ajb2_1069_f7.pdf




Gene trees concordant with species tree
Gene trees with most common conflict to species tree
Gene trees with insufficient sampling to analyze relationship
Remaining conflicting gene trees
Boerhavia torreyana
Boerhavia purpurascens
Boerhavia coccinea
Boerhavia burbidgeana
Okenia hypogea
Boerhavia ciliata
Anulocaulis annulatus
Anulocaulis leiosolenus
Anulocaulis eriosolenus
Nyctaginia capitata
Cyphomeris gypsophiloides
Allionia incarnata
Commicarpus scandens
Mirabilis jalapa
Mirabilis pringlei
Mirabilis multiflora
Abronia glabrifolia
Abronia bigelovii
Abronia carletonii
Abronia fragrans
Abronia latifolia
Abronia maritima
Abronia umbellata
Tripterocalyx carneus
Tripterocalyx crux-maltae
Acleisanthes lanceolata
Acleisanthes purpusiana
Acleisanthes chenopodioides
Acleisanthes obtusa
Bougainvillea stipitata
Bougainvillea spectabilis
Guapira obtusata
Neea psychotrioides
Pisonia aculeata
Pisonia umbellifera
Colignonia ovalifolia
Salpianthus purpurascens
Petiveriaceae Monococcus echinophorus
Petiveriaceae Petiveria alliacea
Petiveriaceae Hilleria latifolia
Petiveriaceae Trichostigma octandrum
0.05
Nyctaginia capitata
Mirabilis multiflora
Abronia umbellata
Pisonia umbellifera