JSE Journal of Systematics and Evolution

doi: 10.1111/jse.12167

Research Article

Using phylogenomics to resolve mega-families: An example from Compositae

Jennifer R. Mandel[1,2*†], Rebecca B. Dikow[3†], and Vicki A. Funk[4]

[1]Department of Biological Sciences, University of Memphis, Memphis, TN 38152, USA
[2]W. Harry Feinstone Center for Genomic Research, University of Memphis, Memphis, TN 38152, USA
[3]Smithsonian Institute for Biodiversity Genomics, Center for Conservation and Evolutionary Genetics, National Zoological Park and Division of Mammals, National Museum of Natural History, Smithsonian Institution, Washington DC 20013-7012, USA
[4]Department of Botany, National Museum of Natural History, Smithsonian Institution, Washington DC 20013-7012, USA
[†]These authors contributed equally to this work.
*Author for correspondence. E-mail: jmandel@memphis.edu. Tel.: 901-678-5130. Fax: 901-678-4457.
Received 17 March 2015; Accepted 25 June 2015; Article first published online 4 August 2015

Abstract Next-generation sequencing and phylogenomics hold great promise for elucidating complex relationships among large plant families. Here, we performed targeted capture of low copy sequences followed by next-generation sequencing on the Illumina platform in the large and diverse angiosperm family Compositae (Asteraceae). The family is monophyletic, based on morphology and molecular data, yet many areas of the phylogeny have unresolved polytomies and interpreting phylogenetic patterns has been historically difficult. In order to outline a method and provide a framework and for future phylogenetic studies in the Compositae, we sequenced 23 taxa from across the family in which the relationships were well established as well as a member of the sister family Calyceraceae. We generated nuclear data from 795 loci and assembled chloroplast genomes from off-target capture reads enabling the comparison of nuclear and chloroplast genomes for phylogenetic analyses. We also analyzed multi-copy nuclear genes in our data set using a clustering method during orthology detection, and we applied a network approach to these clusters—analyzing all related locus copies. Using these data, we produced hypotheses of phylogenetic relationships employing both a conservative (restricted to only loci with one copy per targeted locus) and a multigene approach (including all copies per targeted locus). The methods and bioinformatics workflow presented here provide a solid foundation for future work aimed at understanding gene family evolution in the Compositae as well as providing a model for phylogenomic analyses in other plant mega-families.

Key words: chloroplast, gene-tree, network, next-generation sequencing, nuclear, phylogenetics.

Families of flowering plants fall easily into artificial groups based on numbers of accepted, extant species. At the extreme are the mega-families, based on the Angiosperm Phylogeny Website (Stevens, 2001 onward): Compositae (ca. 24 000 species with as many as 35 000 estimated), Orchidaceae (ca. 22 000 species with as many as 28 000 estimated), and Leguminosae (ca. 19 500 species). Estimates of the number of flowering plant species vary greatly but most are between 250 000 and 350 000, so 19%–26% of flowering plant species are in these three mega-families. These families are not closely related to one another; they are found in the three major clades within the Angiosperms (asterids, monocots, rosids, respectively) and they frequently have very different habits. Mega-families present special sets of challenges, not the least of which is that often various clades within the family are studied by different individuals often using different methods, genes, or gene regions in molecular research projects. Integrating these studies is not trivial and makes it difficult to work across groups to study evolution at the family level.

Advances in next-generation sequencing (NGS) have given researchers the ability to produce massive amounts of genomic data that allow us to efficiently assay hundreds of loci across many taxa, and these data can be used to resolve relationships at the species-level, as well as among major clades (e.g., Jarvis et al., 2014; Misof et al., 2014). When obtaining fresh RNA (for phylogenomic approaches using transcriptome sequencing; e.g., Chan & Ragan, 2013) is not feasible as for museum or herbarium specimens, next-generation methods that make use of genomic DNA are necessary. Recent work in a number of taxonomic groups has utilized DNA sequence capture (target enrichment plus NGS) of conserved sequences for phylogenomic analyses (e.g., Faircloth et al., 2012; Mandel et al., 2014; Weitemier et al., 2014). Phylogenetic studies employing NGS have provided much insight into important evolutionary questions within these groups related to the timing and pattern of speciation events and elucidating relationships among key lineages (Liu et al., 2015).

To begin to answer these and other questions in the largest flowering plant family, Compositae, we have designed

sequence capture probes targeting a conserved ortholog set (COS; a set of putatively single/low copy homologous sequences) for understanding relationships among members. These COS probes were developed by identifying homologues to *Arabidopsis* single copy genes in three Compositae (safflower, sunflower, lettuce; see Mandel et al., 2014 for a detailed description of the development). Because the sunflower genome is more than 81% repetitive (Staton et al., 2012), much of its genome (likely also for many other species of Compositae) is not amenable to phylogenetic analysis, gene target enrichment is one way to capture only potentially phylogenetically informative regions of the genome. Additionally, as members of Compositae exhibit ploidy levels that vary between 2× and 48× (Semple & Watanabe, 2009), gene copy number causes complications in all possible NGS approaches. Because paralogs are gene copies that have, by definition, duplicated since the last split with a common ancestor, they do not necessarily follow the "species" history that one looks to hypothesize with a phylogeny. We sought to capture single or low copy genes to minimize this challenge and chose bioinformatics methods that also minimize the paralogy problem. In addition, we explore other options that consider multiple gene copies, as massive duplications have been a hallmark of the evolutionary history of flowering plants, and understanding the full extent of this history and diversification requires an exploration of genes that have been duplicated many times across the Compositae tree.

An estimated 7%–10% of all flowering plants belong to the Compositae mega-family (Fig. 1 depicts some of the phenotypic variation found in the family). Members of the family are widespread and occur throughout the world, but are most abundant in open areas with seasonal climates such as Mediterranean regions, deserts, prairies, steppes, and mountains. While many species have restricted ranges in geographic areas that are threatened with high extinction rates (i.e., Pacific Islands, Cape Floristic Region, páramo and puna regions of the Andes), the family also includes some of the world's most noxious weeds (i.e., dandelion, ragweed, thistle) some of which cost the United States roughly 35 billion USD annually (Pimentel et al., 2005). Numerous species produce novel secondary compounds that have many industrial and biomedical uses such as artificial sweeteners (*Stevia*), insecticides (pyrethrins from *Chrysanthemum*), rubber (*Parthenium*), and the anti-malarial compound artemisinin from *Artemisia anuua* (Eckstein-Ludwig et al., 2003), and more than 40 species have been domesticated for food and medicinal uses (i.e., artichokes, *Echinacea*, endive, lettuce, safflower, sunflower, tarragon, and the popular and interesting wormwood used to give the liquor absinthe its distinctive character). There are also many important garden ornamentals, such as ageratums, asters, chrysanthemums, cosmos, dahlias, marigolds, and zinnias.

The Compositae family is monophyletic based on morphology as well as molecular data, but resolving the relationships among major groups has always been difficult. The most comprehensive phylogeny to date is a meta-tree (Funk & Specht, 2007) that was constructed using a base tree of 10 chloroplast loci (Funk et al., 2009a; based on: Panero & Funk, 2008; Funk & Chan, 2009; Pelser & Watson, 2009; Baldwin, 2009); this meta-tree included 900 of the 1700 genera found in the family. Several areas of the phylogeny have unresolved polytomies involving taxa that vary in key morphological traits; thus, well-supported hypotheses of character evolution cannot be developed (e.g., Ortiz et al., 2009). Compositae are an excellent group for studying biogeography (Funk et al., 2009b), pollen evolution, secondary chemistry, paleopolyploidy (Barker et al., 2008), domestication (Dempewolf et al., 2008), and invasions (Lai et al., 2012). All of these studies are hampered by the unresolved areas of the base tree. Mandel et al. (2014) previously demonstrated the utility of sequence capture and NGS for resolving relationships within the family. Here we expand upon that previous work by considering multi-copy genes in our data set using a clustering approach, during orthology detection prior to gene-tree phylogenetic analysis. Instead of excluding taxa with greater than one match for a particular locus (as in Mandel et al., 2014), all copies are analyzed by clustering these into "genes within genes." Using this output, a network approach is applied to begin to identify and understand gene tree incongruence and the complex dynamics of plant nuclear gene families. We also expand our taxon sampling to include additional species with close tribal relationships to apply our gene clustering and network workflow at both broad and fine evolutionary scales. Finally, we have assembled chloroplast genomes from off-target capture reads, thus enabling the comparison of both the biparentally-inherited nuclear and maternally-inherited chloroplast genomes for phylogenetic analyses. There have been methods proposed to identify orthologous copies of genes in a phylogenetic context (e.g., Dunn et al., 2013; Kocot et al., 2013; Yang & Smith, 2014) as inclusion of paralogs in phylogenetic analyses can be misleading. Proliferation of gene copies in plants has often led to the diversification of life histories and morphologies (Soltis et al., 2009), and Compositae is a family of plants within which there are dramatic variations in ploidy, and thus gene copy number. The two approaches presented here acknowledge both of these points: the conservative approach rejects all multiple copies; and the multigene approach attempts to consider all copies, which will help us understand character evolution and diversification.

Material and Methods

Taxon sampling

The motivation for choosing taxa for this study was to investigate and reconstruct both broad- and fine-scale phylogenetic relationships within Compositae. Twenty-three taxa from nine clades (tribes and subtribes) were selected including species that span the entire family and its sister group, the Calyceraceae (Table 1). Ten of the species are distributed across the phylogeny in an effort to test the usefulness of our approach in all areas in the family. Seven species are from the tribe Heliantheae subtribe Ecliptinae, and consist of two related genera from one of the most diverse radiations of Compositae in the Hawaiian Islands, *Lipochaeta* and *Melanthera* (Funk et al., 2009a). This clade was selected because it contains species that are diploids as well as species that are tetraploids. We selected the final six species from the tribe Heliantheae subtribe Helianthinae consisting of representatives from the genus *Helianthus* (including three rare/endangered taxa and the economically important

Fig. 1. Diversity of the Compositae. **A,** *Chuquiraga* (Barnadesieae-Barnadesoideae), a representative of the clade that is the sister group of the rest of the family. Note the paired axillary spines and brightly colored bracts surrounding the flowers. **B,** Two members of different tribes gowing together: left = *Weneria* (ca. 1 cm tall; Senecioneae-Asteroideae) and right = *Perezia* (Nassauvieae-Mutisioideae). The inset is the collecting locality, near Lago Chungara and Volcán Parinacota, Chile. **C,** *Melanthera* (Heliantheae-Asteroideae), growing on lava near the ocean on Kaena Point on the northeast tip of Oahu, Hawaii. **D,** *Arctotis* (Arctotideae-Cichoroideae), from the Western Cape of South Africa. **E,** *Carduus* (Cardueae-Carduoideae), a thistle that is often weedy in many parts of the world including this one in Patagonia, Argentina. Note the spiny bracts and the brightly colored flower parts. **F,** *Xenophyllum,* growing on bare rocky slopes at nearly 5000 m in elevation, Chile. The inset is an enlargement of the flowering head showing dark purple, fused bracts, thick fleshy leaves, and small yellow flowers. (A and E, photo by JM Bonifacino, remainder by VA Funk)

domesticated sunflower) and its sister genus, *Phoebanthus*. These three groups of species should allow us to examine the utility of the methods across the family, between diploids and polyploids, and in a genus of closely related species.

Hybridization, sequencing, and data processing

Genomic DNA of each species was extracted using the DNeasy plant mini kit (Qiagen, Valencia, California, USA) and a barcoded sequencing library was constructed using either the TruSeq DNA Sample Preparation kit (Illumina, San Diego, California, USA) or the NEBNext Ultra DNA Library Prep Kit (New England Biolabs, Ipswich, Massachusetts, USA). Sequence capture was performed using a custom probe set, MyBaits designed by MYcroarray (Ann Arbor, Michigan, USA) and described in Mandel et al. (2014). DNA libraries were checked for quality using 2100 Bioanalyzer (Agilent, Santa Clara, California, USA), quantified using a Qubit 2.0 Fluorometer (Life Technologies, Grand Island, New York, USA), pooled, and sequenced on either two lanes of an Illumina HiSeq 2000 sequencer (2×100 bp read length) or one run of the Illumina MiSeq sequencer (2×250 bp read length).

Table 1 Species sampled and collection information

Family or tribe	Genus	Species	Authority	Location	Date	Collector(s)	#
Calyceraceae	*Nastanthus*	*patagonicus*[+]	Spreg.	Argentina: Santa Cruz, Rio Chico	14-Dec-2009	Bonifacino, M. & Funk, V.A.	4016[†]
Barnadesieae	*Fulcaldea*	*stuessyi*[+]	Roque & V.A. Funk	Brazil: Bahia	09-Aug-2010	Abreu, I.S.	123[†]
Mutisieae	*Gerbera*	*hybrida*[+]	n/a	Greenhouse grown cutting, Terra Nigra USA	22-Oct-2013	Mandel, J.R.	105[‡]
Cardueae	*Carthamus*	*tinctorius*[+]	L.	Voucher n/a, USDA, PI 592391	n/a	n/a	n/a
Cichorieae	*Taraxacum*	*kok-saghyz*[+]	Rodin	Greenhouse grown seed[≠] USDA, W6 35156	27-Aug-13	Mandel, J.R.	102[‡]
Vernonieae	*Centrapalus*	*pauciflorus*[+]	(Willd.) H. Rob.	Greenhouse grown seed USDA, PI 312852	22-Oct-2013	Mandel, J.R.	104[‡]
Senecioneae	*Senecio*	*vulgaris*[+]	L.	Washington DC: NMNH	07-Nov-2011	Funk, V.A.	12774[†]
Gnaphalieae	*Pseudognaphalium*	*obtusifolium*[+]	(L.) Hilliard & B.L. Burtt	VA: Fairfax Co., Falls Church	12-Sep-2011	Funk, V.A.	12773[†]
Gnaphalieae	*Antennaria*	*pulcherrima*	(Hook.) Greene	Gothic, CO	10-Jul-1991	Bayer, R.J., Minish & Francis	CO-91012[§]
Eupatorieae	*Conoclinium*	*coelestinum*[+]	(L.) DC.	VA: Fairfax Co., Falls Church	11-Sep-2001	Funk, V.A.	12769[†]
Heliantheae	*Melanthera*	*micrantha*	(Nutt.) W. L. Wagner & H. Rob.	NTBG[¶], Kauai		Keeley, S.	
Heliantheae	*Melanthera*	*venosa*	(Sherff) W. L. Wagner & H. Rob.	Nohonohoe, Hawaii		Keeley, S.	
Heliantheae	*Melanthera*	*faurieri*	(H. Lév.) W. L. Wagner & H. Rob.	Haeleele Valley, Kauai		Keeley, S.	
Heliantheae	*Melanthera*	*subcordata*	(A. Gray) W. L. Wagner	Waimea Falls Park, Kauai	27-Apr-1992	Keeley, S.	[†]
Heliantheae	*Melanthera*	*biflora*	(L.) Wild	NTBG[¶], Okinawa Island	27-Apr-1992	Keeley, S.	[†]
Heliantheae	*Lipochaeta*	*lobata*	(Gaud.) DC.	Hanaula Rd., Maui	01-Mar-1993	Keeley, S.	[†]
Heliantheae	*Lipochaeta*	*connata*	(Gaud.) DC.	Waimea Canyon, Kauai		Keeley, S.	[†]
Heliantheae	*Phoebanthus*	*tenuifolius*[+]	S.F. Blake	Greenhouse grown seed collected FL: Liberty Co.	10-Sep-2010	Mason, C.M	101[‡]
Heliantheae	*Helianthus*	*porteri*[+]	(A. Gray) Pruski	Greenhouse grown seed collected GA: DeKalb Co.[≠]	22-Oct-2013	Mandel, J.R.	103[‡]
Heliantheae	*Helianthus*	*verticillatus*[+]	Small	Greenhouse grown seed collected TN: Madison Co.	01-Sep-2004	Mandel, J.R.	101[‡]
Heliantheae	*Helianthus*	*niveus* ssp. *tephrodes*[+]	(A. Gray) Heiser	Voucher n/a, USDA, PI 613758	n/a	n/a	n/a
Heliantheae	*Helianthus*	*argophyllus*[+]	Torr. & A. Gray	Voucher n/a, USDA, PI 435623	n/a	n/a	n/a
Heliantheae	*Helianthus*	*annuus*[+]	L.	Voucher n/a, USDA, PI 603989	n/a	n/a	n/a

#Voucher number. [†]US Herbarium, [‡]GA Herbarium, [§]ALTA Herbarium, [¶]NTBG National Tropical Botanical Garden. [≠]Voucher specimen is individual from same population. DNA from some taxa studied had been extracted for other projects and the plants were not vouchered, they are listed as n/a. [+]Sequence published in Mandel et al., 2014.

The resulting data were processed following the bioinformatic and phylogenetic workflow described in Mandel et al. (2014) using the custom scripts and publicly available code that can be found at https://github.com/Smithsonian/Compositae-COS-workflow. Briefly, raw sequence reads were quality trimmed, converted from FASTQ to FASTA, and either subjected to a BLAST filtering step to screen for reads that contained DNA of targeted loci and assembly (HiSeq data) or sent straight to assembly (MiSeq data). For each taxon, assembly of the cleaned and filtered reads (HiSeq data) was performed using the Velvet *de novo* sequence assembler package (Zerbino & Birney, 2008).

Conservative COS loci

Contigs from each of the 23 species were analyzed in the PHYLUCE pipeline (v. 0.1.0; Faircloth et al., 2012) as described in Mandel et al. (2014). PHYLUCE permits only one contig to match probes from one COS locus and that only probes from one COS locus match one contig. This leads to missing data in the resulting data matrix (see Mandel et al., 2014). While missing data in a phylogenetic context is often considered a short-coming, here, PHYLUCE works to minimize the inclusion of possible paralogs and thus enables us to produce a conservative species tree that follows the history of speciation rather than the signals of duplicated gene copies, which do not necessarily follow the taxon branching pattern. COS loci and copy number have also been depicted by a heatmap created using Gitools (Perez-Llamas & Lopex-Bigas, 2011).

Multigene COS loci

In order to consider multi-copy genes in our dataset, we employed a parallel approach in which we began similarly as above with assembly and the initial LASTZ step of PHYLUCE, but then we associated multiple matches for each COS locus using custom scripts available on GitHub. These COS loci, which often had multiple copies (i.e., multiple assembled contigs with LASTZ hits) for each taxon, were then clustered using USEARCH (Edgar, 2010) at 65% similarity before nucleotide alignment and tree search, allowing each original locus to split into multiple loci. While USEARCH suggests a minimum of 75% similarity for sequence clustering in order to avoid false homology, in our case, we clustered within known COS loci rather than globally and were less concerned with the possibility of false homology. Clusters within COS loci and copy number are depicted by a heatmap created using Gitools (Perez-Llamas & Lopex-Bigas, 2011).

Chloroplast genomes

Even though we did not target the chloroplast genome in these captures, we were able to assemble partial to near complete chloroplast genomes using off-target reads. Reads from captures and WGS sequences (for the taxa sequenced in Mandel et al., 2014) were mapped to the published *Helianthus annuus* chloroplast genome using Geneious R6 and then reads that mapped were *de novo* assembled using Velvet. This process was repeated iteratively to incorporate initially poorly matching reads. Additional chloroplast genomes from Genbank were also included in the final alignment so that the total number of clades was increased to a total of 12 not including the outgroup. Genbank accession numbers for these additional taxa are listed in Table S1.

Nucleotide alignments and phylogenetic analyses

Nucleotide alignments were performed in MAFFT (v. 7.029b; Katoh et al., 2002; as implemented in PHYLUCE for the conservative COS loci dataset, command line for all other datasets). Clusters were aligned separately and treated as separate "loci." Phylogenetic analyses of concatenated data sets were completed under the maximum likelihood optimality criterion in GARLI (v. 2.0; Zwickl, 2006; GTR-gamma model of nucleotide substitution; 100 search replicates; 1000 bootstrap replicates). Conservative COS loci were also analyzed individually as above in GARLI, as were multigene COS loci. Conflict among gene trees is depicted using SuperNetwork functionality of SplitsTree (Huson & Bryant, 2006). ASTRAL (Mirarab et al., 2014), a pseudo-coalescence method, was also used on the conservative and multigene datasets to calculate consensus "species" trees based on individual gene trees.

Results

Raw read information and statistics for all 23 taxa can been found in Table S2. Bioinformatics workflow is shown in Fig. 2. Raw reads have been deposited in the NCBI SRA under BioProject PRJNA236448. A Dryad package (http://dx.doi.org/10.5061/dryad.k9k23) contains all alignments, tree files, and heatmap files referred to below.

Conservative COS loci

PHYLUCE found a total of 871 COS loci for 3 taxa or more and 795 COS loci for 4 taxa or more. Total concatenated alignment length for 871 loci was 274 242 bp (range from 19 769 to 116 769) for each species. Mean COS locus alignment length was 315 bp. Fig. S1 depicts occupancy of these loci for all taxa. This figure is an SVG file, which can be viewed in a web browser or other application that opens SVG files. The Gitools files are also provided on Dryad, which are interactive when viewed within the Gitools software package. Relationships based on phylogenetic analysis of the concatenated dataset with bootstrap numbers above the branches are shown in Fig. 3A and SuperNetwork analysis of gene trees is shown in Fig. 3B. ASTRAL gene tree analysis is shown in Fig. 4.

Multigene COS loci

There were 3997 clusters (which represent all 795 loci) with four or more taxa. Mean alignment length: 818 bp (these alignments were not trimmed as the conservative COS loci were in the PHYLUCE implementation of MAFFT because they were first clustered based on similarity). A heatmap showing the occupancy of the clusters for each species is presented in Fig. S2. A SuperNetwork analysis for sets of these clusters representing low copy number COS loci (one or two clusters per COS locus: 119 total clusters) is shown in Fig. 5A and a SuperNetwork analysis for a random sampling of 100 clusters is shown in Fig. 5B.

Chloroplast genomes

The total chloroplast alignment is 164 333 bp. The phylogenetic tree based on complete and partial chloroplast genomes along with bootstrap support at the nodes (dots indicate

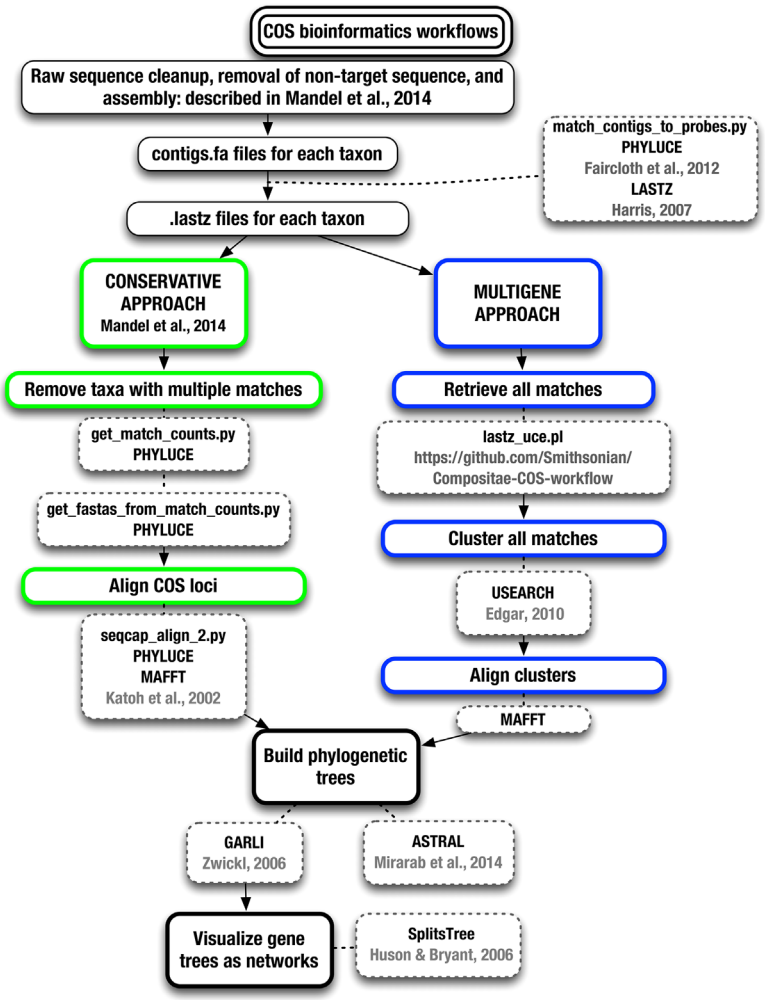

Fig. 2. Bioinformatics workflow.

support 75% or above) is shown in Fig. 6; the bars indicating the percentage of genome coverage relative to the total *Helianthus annuus* chloroplast are shown on the right. This phylogeny contains 38 taxa because an additional 15 samples were added by using chloroplast genomes available from Genbank.

Discussion

We have demonstrated a phylogenomic approach for resolving species relationships in one very large and diverse plant mega-family. This approach includes a targeted enrichment of low copy nuclear genes followed by NGS and an innovative bioinformatics workflow. The bioinformatics workflow takes a parallel and complementary approach to building phylogenetic trees employing a previously published method (Mandel et al., 2014) and a "genes within genes" method presented here. The conservative method rejects putative orthologs with more than one match assuming possible paralogy resulting in a significant fraction of missing data and is indicative of the stringency of this approach. The multigene method makes use of clustering algorithms that group potential orthologs and paralogs and aligns these clusters of genes before phylogenetic reconstruction. Employing this approach allows the analysis of both species- and gene-level network trees providing several schemas for understanding potentially complex evolutionary patterns.

In general the two phylogenies based on concatenated data, nDNA (Fig. 3A) and cpDNA (Fig. 6), agree with topologies found in previous studies using Sanger sequencing (Funk et al., 2009b) but with far better support in the nDNA tree, especially near the base. In Fig. 3A, the two subtribes of the Heliantheae that were included (Ecliptinae and Helianthinae) are sister taxa and the Eupatorieae is sister to those two taxa. The remainder of the subfamily Asteroideae (Gnaphalieae and Senecioneae) are sister to the Heliantheae-Eupatorieae clade. The sister taxon to the Asteroideae is the subfamily Cichorioideae represented by two tribes (Vernonieae and Cichorieae). Below these are the Cardueae, Mutisieae, and Barnadesieae tribes. The phylogenetic hypotheses are further compared below.

Conservative COS loci

As shown in Fig. 3A, the phylogenetic hypothesis based on concatenation of all 795 COS loci shows strong support for all

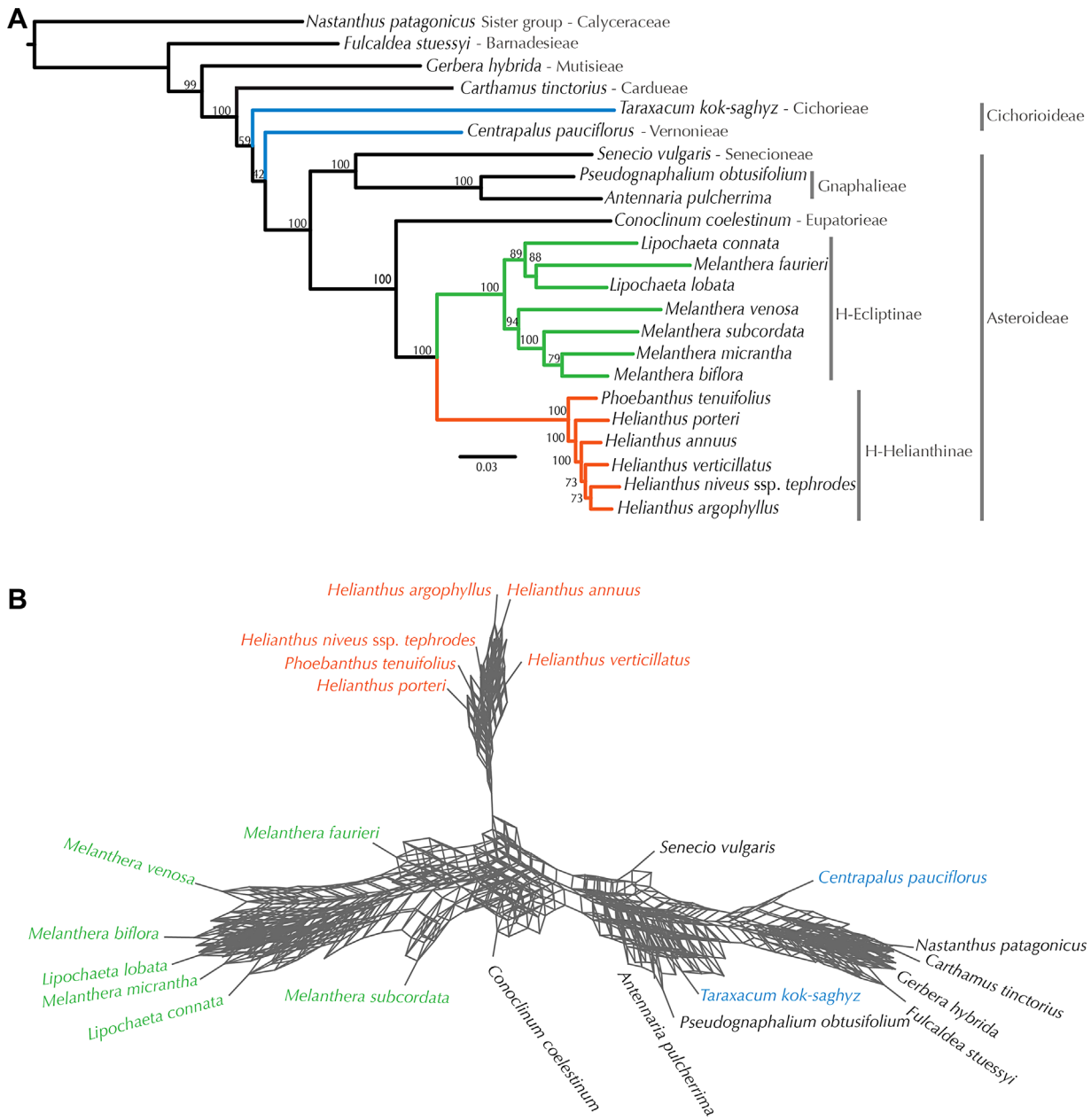

Fig. 3. Phylogeny based on conserved loci. **A,** Relationships based on phylogenetic analysis of conservative COS loci in the concatenated data set. Phylogenetic relationships generated by ML in Garli. Bootstrap values are shown above the branches. H, Heliantheae Alliance. **B,** SuperNetwork for all conservative COS loci. Network of relationships generated in SplitsTree showing gene tree incongruence. Individual gene trees generated with Garli.

nodes except the two involving members of the Cichorioideae, *Taraxacum* and *Centrapalus* (shown in blue). *Centrapalus* is a member of the tribe Vernonieae and *Taraxacum* is a member of the tribe Cichorieae and they are the only members of the subfamily Cichorioideae included in this study. They are expected to be sister taxa as they are in the chloroplast tree (Fig. 6), however, in the nDNA tree (Fig. 3A) they are not sister taxa. This is not a serious problem since they are proximal to one another and the support for grouping *Centrapalus* with the rest of the family is less than 50%. Improving the taxon sampling within these two tribes and the remaining tribes of the subfamily Chicorioideae is expected to resolve this placement. All the other relationships are as expected based on previous hypotheses.

Individual gene tree analyses with Garli are not particularly informative because each "gene," i.e., COS locus, is fairly short (averaging 315 bp) and because there is a large amount of missing data in this matrix due to the conservative orthology detection using PHYLUCE (see occupancy in Fig. S1). PHYLUCE permits only one contig to match probes from one COS locus and only probes from one COS locus match one contig, meaning that loci with multiple copies are removed from the

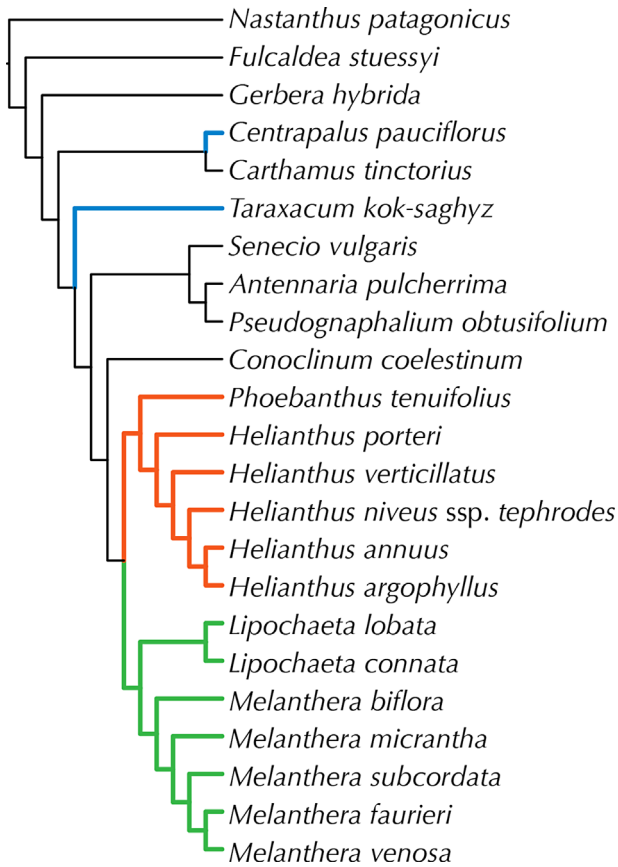

Fig. 4. ASTRAL tree; Conservative COS gene trees analyzed to produce a "species" tree.

final set of orthologs. This is an appropriate approach, when a concatenated species tree (minimizing the chance that paralogy is affecting the phylogenetic relationships) is the goal, and the recovery of topologies supported by previous studies indicates that this approach is successful in this regard.

Two ways we can look at gene-tree incongruence are by analyzing in a pseudo-coalescence framework such as with ASTRAL (Fig. 4) and by a SuperNetwork generated in SplitsTree. A true coalescence method such as that implemented in *BEAST (Heled & Drummond, 2010) is not possible given these data, first because the number of loci is prohibitive and because we do not have each species represented for each gene, a requirement for these methods. The ASTRAL species tree places *Carthamus* sister to *Centrapalus*, which is different from the concatenated hypothesis (Fig. 3A) and also indicates some differences in the *Helianthus* clade (shown in orange in Fig. 3A) and the *Lipochaeta*/*Melanthera* clade (shown in green in Fig. 3A). These more shallow differences are not unexpected, as species of *Helianthus* and *Lipochaeta*/*Melanthera* are thought to have accumulated gene copies throughout the evolution of the family, likely leading to an increased chance of gene tree incongruence due to hidden paralogy. The ASTRAL tree is the only one that finds *Lipochaeta* and *Melanthera* monophyletic, respectively. Further comments on the relationships within these genera are below. The SuperNetwork (Fig. 3B) is an alternative way to look at the gene, or COS, trees that does not require bifurcation. For this network, all 795 COS tree topologies were input into SplitsTree, and in general the major groupings are visible, while the shallower splits are not as discernable. The two taxa indicated in blue, *Centrapalus* and *Taraxacum,* are an exception and are not placed together; additionally, *Carthamus* is placed in different positions according to different analyses. In all three approaches that use the conservative COS loci: concatenation and phylogenetic analysis, gene tree pseudo-coalescence using ASTRAL, and a network using SplitsTree, these three genera have uncertain placement. Further taxon sampling and investigation of gene copy numbers (as discussed below) will elucidate the cause of this uncertainty.

What is missing from this approach, in which only single copy genes are considered, is the ability to understand how the evolution of gene families has impacted both the evolutionary history and functional innovations of Compositae (i.e., pollen presentation; unique primary, secondary, and tertiary head structures; complex secondary chemistry, etc.). By including what look to be multiple copies of the originally designed COS loci (which are putatively single copy in *Arabidopsis*), we can begin to tease apart which copies are orthologous and how our original COS loci may include either gene families or genes within genes for species across Compositae.

Multigene COS loci

As mentioned above, we undertook this approach in order to begin to understand how multiple gene copies can be used to better understand the evolutionary history of Compositae. We first allowed multiple probe matches in LASTZ and then used USEARCH to cluster these initial matches before multiple sequence alignment. Our 795 conservative COS loci, when all probe matches per locus were considered, generated 3997 clusters, which are shown in Fig. S2. From these 3997 alignments, individual gene trees were generated with Garli. This total dataset was a challenge to analyze. Certain species (mostly within *Helianthus*) had up to a hundred or more copies of a particular gene (although this was rare and can be pinpointed to a repetitive section of DNA, COS-924, that showed a bimodal distribution for presence/absence with taxa either having few to no copies or more than 100 copies). These results are interactively viewable in the heatmap (Fig. S2) when visualized in Gitools.

While most species ended up only having a few copies of a particular gene, the fact that this was not uniform across the 3997 clusters (e.g., species A has five copies of cluster 1, one copy of cluster 2; species B has two copies of cluster 1, and six copies of cluster 2; multiplied by 3997 clusters and 23 species), made integrating all 3997 cluster trees into a single diagram uninformative (a star network with no resolution). Finer analysis of groups of these clusters, however, was informative. Fig. 5A is a SuperNetwork based on cluster trees for those clusters for which each species had a maximum of one copy per species (119 clusters). Fig. 5B is a SuperNetwork based on 100 cluster trees selected randomly (without reference to copy number). Fig. 5A is quite similar to Fig. 3B, which is the network based on the conservative COS loci trees. This is not surprising, given that 5A is based on the low copy clusters. Fig. 5A even shows the expected relationship between *Centrapalus* and *Taraxacum*, which is not found in Fig. 3A.

Fig. 5. SuperNeworks. **A,** SuperNetwork for low copy clusters. These 119 are the lowest copy of all 3997 total clusters, having no more than 1 copy. **B,** SuperNetwork for 100 randomclusters.

Fig. 5B, on the other hand, shows species present in multiple parts of the network, indicating more about gene copy evolution than species relationships.

We plan to continue this line of inquiry by looking at the function of the COS loci and grouping the clusters within these by function to analyze as functional units. One hypothesis is that we might expect loci with similar or coordinated function to show similar patterns of duplication and diversity throughout the family. Additionally, copies may be associated with ploidy level and this possibility will also be investigated.

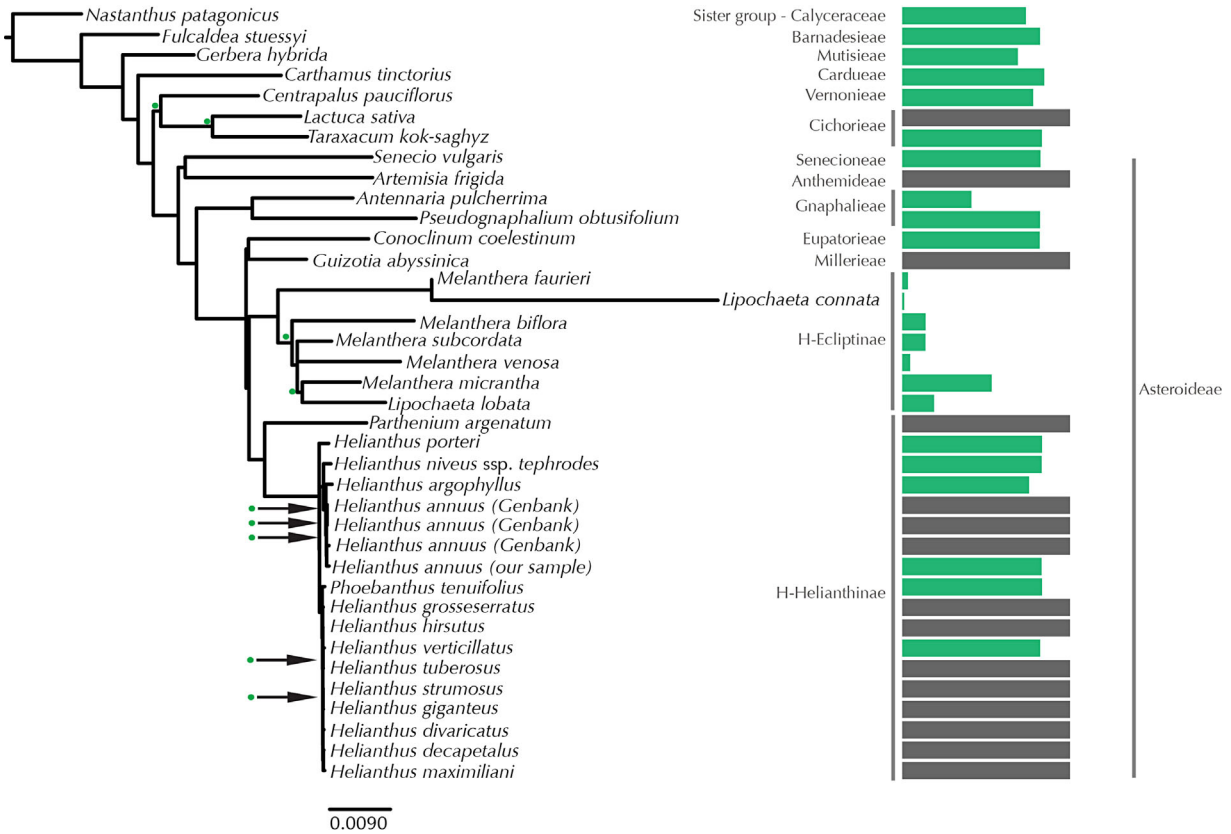

Fig. 6. Chloroplast genome phylogenetic hypothesis. Phylogenetic relationships generated by ML in Garli. Bootstrap values are shown above the branches and bars (green for taxa sequenced in this study, black for sequences downloaded from GenBank) indicate chloroplast genome coverage. H, Heliantheae Alliance.

Chloroplast genomes

The chloroplast tree (Fig. 6) is congruent with the conservative COS tree in many respects, except for the placement of *Taraxacum* which is now found in a clade with *Centrapalus*. The addition of *Lactuca* is likely partly responsible since it is in the same tribe as *Taraxacum*. Further taxon sampling within Cichorieae and Vernonieae is necessary to better resolve this part of the tree. Those nodes on the conservative COS tree had the lowest support, while they have some of the highest support on the chloroplast tree. In general, the support values on the chloroplast tree are lower than the conservative COS tree, likely because the number of parsimony informative sites is less than half for the chloroplast dataset (8458/164 333 sites or 5% for the chloroplast as compared to 30 873/274 242 or 11% for the conservative COS dataset). While we did not use parsimony as an optimality criterion when building the phylogeny, calculating the number of parsimony informative sites remains a rough guide to the phylogenetic information content in a given dataset. Those taxa with fewest data (the *Lipochaeta*/*Melanthera* complex) were sequenced on a MiSeq rather than a HiSeq, as all other taxa were sequenced. The MiSeq generates longer, but fewer, reads per taxon and this had a significant effect on the chloroplast assemblies. For these taxa, we also did not have WGS reads to complement the enrichments, leading to fewer total reads incorporated in the chloroplast assemblies. The long branch leading to *L. connata* may be due to the lack of data for that taxon. Finally, it is worth noting that in addition to off-target chloroplast DNA present in our sequencing, other high copy genomic is also present in our data set, e.g., ribosomal genes, mitochondrial DNA, that could be useful for phylogenetic analyses in the future.

Taxon placement in a "difficult" complex

The *Lipochaeta*/*Melanthera*/*Wedelia* complex has a complicated history because there is a dearth of consistent characters to use for descriptions and keys. For details see discussions in Gardner (1979), Wagner & Robinson (2001), and Orchard (2013). The results of this study shed light on some of the relationships within this group and provide information for future taxon sampling efforts in this difficult clade. However there are several important points that are relevant to the discussion. First, *Lipochaeta* and *Melanthera* have long been thought to be closely related to one another and to *Wedelia* and they were placed in the same subtribe (Stuessy, 1977). Second, although the Hawaiian *Lipochaeta* contained both diploid (five-merous corollas) and allotetraploid (four-merous corollas) species, most treatments placed them in a single genus, *Lipochaeta,* until 2001 when Wagner & Robinson transferred the diploids into the genus *Melanthera* (a pantropical sometimes weedy genus). Third, the allotetraploid

Hawaiian species (base chromosome number of $x = 26$) were thought to be the result of a hybridization event between a *Wedelia* ($x = 11$) and a diploid *Lipochaeta* (base chromosome number of $x = 15$). Finally, there is an as yet unpublished molecular study that shows all of the Hawaiian species as part of a single lineage and the tetraploids nested within the diploids (Funk, unpublished data). There are two differences in the *Lipochaeta*/*Melanthera* clade in phylogenies found in Figs. 3 (nDNA phylogeny) and 5 (cpDNA phylogeny): Fig. 3 places the two tetraploid *Lipochaeta* species included in the study together in the same clade but with a diploid (ploidy inferred from corolla) species, *M. fauriei,* while in the cpDNA tree one of the *Lipochaeta* species (*L. lobada*) moves to the other clade. The other difference is in the position of *M. biflora* which is highly nested in one of the clades in the nDNA tree and the first branch of the same clade in the cpDNA tree. These two differences probably have different causes. For the placement of *M. fauriei* it seems that this may be the result of fewer data and/or a possible misidentification of the voucher (Fig. 6). Concerning the placement of *M. biflora*, the different positions may be due to the fact that the tetraploids are of hybrid origin, and we have not yet identified the diploid parent—either *M. biflora or M. faurieri* is a possibility. It is interesting that Fig. 4, the ASTRAL tree, formed from Conservative COS trees analyzed to produce a "species" tree, show both the diploid and tetraploid species in monophyletic groups that are sister taxa.

What is very positive in these results is the high support values found in the phylogenies, especially the nDNA tree, far higher than is generally found for island clades where rapid radiation is the norm. We believe our approach holds great promise especially after adding to the analysis the remaining extant taxa in this group to the analysis.

Outlook

The work presented here is a first step toward tackling phylogenomic-level questions in mega-families. In the Compositae, we are working toward generating a base tree (backbone tree) with at least 200 taxa in order to establish a framework for future evolutionary studies. A panel of experts is being formed that will help train molecular systematists throughout the family to use the same methods and produce phylogenies for each of the tribes. These can then be studied in conjunction with the base tree. The production of a robust, well-sampled phylogeny will allow us to study traits that have led to the remarkable phenotypic diversity and evolutionary success of this family. Furthermore, genome-level phylogenetic questions can be addressed including (i) whether function/annotation of genes can inform phylogenetic analyses, (ii) how many loci across different taxonomic scales are necessary to provide robust trees, and (iii) will these approaches be able to scale up to hundreds or thousands of taxa. Phylogenomic approaches, such as those presented here, hold great promise for other large and difficult families, not just in plants but across the tree of life. Still, factors such as missing data, paralogy, polyploidy, and sequence misalignment may present special problems for phylogenomic analyses, and novel algorithms and programs that utilize genome-level sequence information for handling these types of issues will be critical for moving the field forward.

Acknowledgements

The authors thank the Undersecretary for Science, Smithsonian Institution, for the Next Generation Sequencing Small Grant to VAF. This research was also supported in part by the W. Harry Feinstone Center for Genomic Research, University of Memphis. The authors also thank Randall Bayer, John Burke, Brant Faircloth, Travis Glenn, Rishi Masalia, and Loren Rieseberg for helpful discussion. Sterling Keeley (UH) and Randall Bayer kindly sent DNA of the *Lipochaeta*/*Melanthera* complex and *Antennaria* (respectively). Portions of the computational work were conducted in and with the support of the L.A.B. facilities of the National Museum of Natural History and the newly organized Smithsonian Institute for Biodiversity Genomics.

References

Baldwin BG. 2009. Heliantheae alliance. In: Funk VA, Susanna A, Stuessy T, Bayer R eds. *Systematics, evolution, and biogeography of Compositae.* Vienna: IAPT. 689–730.

Barker MS, Kane NC, Kozik A, Michelmore RW, Matvienko M, Knapp SJ, Rieseberg LH. 2008. Multiple paleopolyploidizations during the evolution of the Asteraceae reveal parallel patterns of duplicate gene retention after millions of years. *Molecular Biology and Evolution* 25: 2445–2455.

Chan CX, Ragan MA. 2013. Next-generation phylogenomics. *Biology Direct* 8: 3.

Dempewolf H, Rieseberg LH, Cronk Q. 2008. Crop domestication in the Compositae: A family-wide trait assessment. *Genetic Resources and Crop Evolution* 55: 1141–1157.

Dunn C, Howison M, Zapata F. 2013. Agalma: An automated phylogenomics workflow. *BMC Bioinformatics* 14: 330.

Eckstein-Ludwig U, Webb RJ, Van Goethem ID, East JM, Lee AG, Kimura M, O'Neill PM, Bray PG, Ward SA, Krishna S. 2003. Artemisinins target the SERCA of *Plasmodium falciparum*. *Nature* 424: 957–961.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26: 2460–2461.

Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology* 61: 717–726.

Funk VA, Anderberg AA, Baldwin BG, Bayer RJ, Bonifacino JM, Breitwieser I, Brouillet L, Carbajal R, Chan R, Coutinho AXP, Crawford DJ, Crisci JV, Dillon MO, Freire SE, Galbany-Casals M, Garcia-Jacas N, Gemeinholzer B, Gruenstaeudl M, Hansen HV, Himmelreich S, Kadereit JW, Källersjö M, Karaman-Castro V, Karis PO, Katinas L, Keeley SC, Kilian N, Kimball RT, Lowrey TK, Lundberg J, McKenzie RJ, Tadesse M, Mort ME, Nordenstam B, Oberprieler C, Ortiz S, Pelser PB, Randle CP, Robinson H, Roque N, Sancho G, Semple JC, Serrano M, Stuessy TF, Susanna A, Unwin M, Urbatsch L, Urtubey E, Vallès J, Vogt R, Wagstaff S, Ward J, Watson LE. 2009a. Compositae metatrees: The next generation. In: Funk VA, Susanna A, Stuessy T, Bayer RJ eds. *Systematics, evolution, and biogeography of Compositae.* Vienna: IAPT. 747–777.

Funk VA, Chan R. 2009. Introduction to the Cichoroideae. In: Funk VA, Susanna A, Stuessy T, Bayer RJ eds. *Systematics, evolution, and biogeography of Compositae.* Vienna: IAPT. 335–342.

Funk VA, Specht C. 2007. Meta–trees: Grafting for a global perspective. Proceedings of the Biological Society of Washington 120: 232–240.

Funk VA, Susanna A, Stuessy T, Bayer RJ eds. 2009b. *Systematics, evolution, and biogeography of the Compositae*. Vienna: IAPT.

Gardner RC. 1979. Revision of *Lipochaeta* (Compositae: Heliantheae) of the Hawaiian Islands. *Rhodora* 81: 291–343.

Harris RS. 2007. *Improved pairwise alignment of genomic DNA*. Ph.D. Thesis. Pennsylvania: The Pennsylvania State University.

Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution* 27: 570–580.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23: 254–267.

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* 346: 1320–1331.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. Mafft: A novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acid Research* 30: 3059–3066.

Kocot KM, Citarella MR, Moroz LL, Halanych KM. 2013. PhyloTree-Pruner: A phylogenetic tree-based approach for selection of orthologous sequences for phylogenomics. *Evolutionary Bioinformatics Online* 9: 429.

Lai Z, Kane NC, Kozik A, Hodgins KA, Dlugosch KM, Barker MS, Matvienko M, Yu Q, Turner KG, Pearl SA, Bell GDM, Zou Yi, Grassa C, Guggisberg A, Adams KL, Anderson JV, Horvath DP, Kesseli RV, Burke JM, Michelmore RW, Rieseberg LH. 2012. Genomics of Compositae weeds: EST libraries, microarrays, and evidence of introgression. *American Journal of Botany* 99: 209–218.

Liu L, Xi Z, Wu S, Davis C, Edwards SV. 2015. Estimating phylogenetic trees from genome-scale data. *arXiv*:1501.03578.

Mandel JR, Dikow RB, Funk VA, Masalia RR, Staton SE, Kozik A, Michelmore RW, Rieseberg LH, Burke JM. 2014. A target enrichment method for gathering phylogenetic information from hundreds of loci: An example from the Compositae. *Applications in Plant Sciences* 2: 1300085.

Mirarab S, Reaz R, Bayzid Md S, Zimmermann T, Swenson MS, Warnow T. 2014. ASTRAL: Genome-scale coalescent-based species tree estimation. *Bioinformatics* 30: i541–i548.

Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C. 2014. Phylogenomics resolves the timing and pattern of insect evolution. *Science* 346: 763.

Orchard AE. 2013. The *Wollastonia*/*Melanthera*/*Wedelia* generic complex (Asteraceae: Ecliptinae), with particular reference to Australia and Malesia. *Nuytsia* 23: 337–466.

Ortiz S, Bonifacino JM, Crisci JV, Funk VA, Hansen HV, Hind DJN, Katinas L, Roque N, Sancho G, Susanna A, Telleria MC. 2009. The basal grade of Compositae: Mutisieae (sensu Cabrerra). In: Funk VA, Susanna A, Stuessy T, Bayer RJ eds. *Systematics, evolution, and biogeography of Compositae*. Vienna: IAPT. 747–777.

Panero JL, Funk VA. 2008. The value of sampling anomalous taxa in phylogenetic studies: Major clades of the Asteraceae revealed. *Molecular Phylogenetics and Evolution* 47: 757–782.

Pelser PB, Watson L. 2009. Introduction to Asteroideae. In: Funk VA, Susanna A, Stuessy T, Bayer RJ eds. *Systematics, evolution, and biogeography of Compositae*. Vienna: IAPT. 495–502.

Perez-Llamas C, Lopez-Bigas N. 2011. Gitools: Analysis and visualisation of genomic data using interactive heat-maps. *PLoS ONE* 6: e19541.

Pimentel D, Zuniga R, Morrison D. 2005. Update on the environmental and economic costs associated with alien-invasive species in the United States. *Ecological Economics* 52: 273–288.

Semple JC, Watanabe K. 2009. A review of chromosome numbers in Asteraceae with hypotehses on chromosomal base number evolution. In: Funk VA, Susanna A, Stuessy T, Bayer RJ eds. *Systematics, evolution, and biogeography of Compositae*. Vienna: IAPT. 61–72.

Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng CF, Sankoff D, Depamphilis CW, Wall PK, Soltis PS. 2009. Polyploidy and angiosperm diversification. *American Journal of Botany* 96: 336–348.

Staton SE, Hartman Bakken B, Blackman BK, Chapman MA, Kane NC, Tang S, Ungerer MC, Knapp SJ, Rieseberg LH, Burke JM. 2012. The sunflower (*Helianthus annuus* L.) genome reflects a recent history of biased accumulation of transposable elements. *Plant Journal* 72: 142–153.

Stevens PF. 2001 onwards. Angiosperm Phylogeny Website. Version 12, July 2012 [continuously updated]. http://www.mobot.org/MOBOT/research/APweb/.

Stuessy TF. 1977. Heliantheae—systematic review. In: Harborne JB, Heywood VH, Turner BL eds. *The biology and chemistry of the Compositae*. New York: Academic Press. 621–671.

Wagner WL, Robinson H. 2001. *Lipochaeta* and *Melanthera* (Asteraceae: Heliantheae subtribe Ecliptinae): Establishing their natural limits and a synopsis. *Brittonia* 53: 539–561.

Weitemier K, Straub SCK, Cronn RC, Fishbein M, Schmickl R, McDonnell A, Liston A. 2014. Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences* 2(9): 1400042.

Yang Y, Smith SA. 2014. Orthology inference in non-model organisms using transcriptomes and low coverage genomes: Improving accuracy and matrix occupancy for phylogenomics. *Molecular Biology and Evolution* 31(11): 3081–3092.

Zerbino DR, Birney E. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* 18: 821–829.

Zwickl DJ. 2006. *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion*. Ph.D. Thesis. Austin: University of Texas at Austin.

Supplementary Material

The following supplementary material is available online for this article at http://onlinelibrary.wiley.com/doi/10.1111/jse.12167/suppinfo:

Fig. S1. Heatmap for conservative COS.

Fig. S2. Heatmap for multigene COS.

Table. S1. GenBank taxa names and accession identifiers.

Table. S2. Next-gen sequencing statistics.