Most Compositae are descendants of a paleohexaploid and all share a paleotetraploid ancestor with the Calyceraceae

Michael S. Barker[2,6], Zheng Li[2], Thomas I. Kidder[2], Chris R. Reardon[2], Zhao Lai[3], Luiz O. Oliveira[4], Moira Scascitelli[5], and Loren H. Rieseberg[3,5]

[2]Department of Ecology & Evolutionary Biology, University of Arizona, Tucson, AZ 85721 USA

[3]Department of Biology and Center for Genomics and Bioinformatics, Indiana University, Bloomington, IN 47405 USA

[4]Departamento de Bioquímica e Biologia Molecular, Universidade Federal de Viçosa, Viçosa, Brazil

[5]Department of Botany and Biodiversity Research Centre, University of British Columbia, Vancouver, BC, Canada

[6]Corresponding author: msbarker@email.arizona.edu

Running title: Nested paleopolyploidies in the Compositae and Calyceraceae

[1]Manuscript received _____; revision accepted _____.

ABSTRACT

Premise of the study: Like many other flowering plants, members of the Compositae (Asteraceae) have a polyploid ancestry. Previous analyses have found evidence for an ancient duplication or possibly triplication in the early evolutionary history of the family. We sought to better place this paleopolyploidy in the phylogeny and assess its nature.

Methods: We sequenced new transcriptomes for *Barnadesia*, the lineage sister to all other Compositae, and four representatives of closely related families. Using a recently developed algorithm, MAPS, we analyzed nuclear gene family phylogenies for evidence of paleopolyploidy.

Key results: We found that the previously recognized Compositae paleopolyploidy is also in the ancestry of the Calyceraceae. Our phylogenomic analyses uncovered evidence for a successive second round of genome duplication among all sampled Compositae except *Barnadesia*.

Conclusions: Our analyses of new samples with new tools provide a revised view of paleopolyploidy in the Compositae. Together with results from recent linkage maps, our results suggest that the Compositae and Calyceraceae have a common paleotetraploid ancestor and most Compositae are descendants of a paleohexaploid. Although paleohexaploids have been previously identified, this is the first example where the paleotetraploid and paleohexaploid lineages have survived over tens of millions of years. The complex polyploidy in the ancestry of the Compositae and Calyceraceae represents a unique opportunity to study the long-term evolutionary fates and consequences of different ploidal levels.

INTRODUCTION

Polyploidy, or speciation by whole genome duplication (WGD), is widespread throughout the history of vascular plants. Nearly 15% of flowering plant speciation events are caused by polyploidy (Wood et al., 2009) and approximately 30% of plant species are recent polyploids (Wood et al., 2009; Mayrose et al., 2011; Barker et al., 2015). Numerous paleopolyploidies have been described among and within families (Schlueter et al., 2004; Cui et al., 2006; Jaillon et al., 2007; Barker et al., 2008, 2009; Rensing et al., 2008; Tang et al., 2008; Soltis et al., 2009; Schmutz et al., 2010; Shi et al., 2010; Velasco et al., 2010; Wang et al., 2011; Arrigo and Barker, 2012; Jiao et al., 2012, 2014; Tomato Genome Consortium, 2012; Vekemans et al., 2012; Cheng et al., 2013; Ibarra-Laclette et al., 2013; Ming et al., 2013; Cai et al., 2014; Estep et al., 2014; Sveinsson et al., 2014; Cannon et al., 2015; Edger et al., 2015; Li et al., 2015; Marques et al., 2015; Yang et al., 2015), as well as in the ancestry of all seed plants (Jiao et al., 2011; Li et al., 2015). Although we know the approximate phylogenetic placement of many paleopolyploidies, we often understand less about their nature. The formation of polyploid species may involve duplication of the genome(s) from one or more species. The outcomes include autopolyploids, allopolyploids, and a series of intermediates such as segmental allopolyploids (Kihara and Ono, 1926; deWet, 1980; Ramsey and Schemske, 1998; Barker et al., 2015). Crosses among taxa of different ploidal levels may lead to hexaploids and other higher level polyploids. Given the potential for polyploids of various natures to have different evolutionary outcomes (Otto and Whitton, 2000; Doyle et al., 2008; Hegarty and Hiscock, 2008; Soltis et al., 2010; Arrigo and Barker, 2012; Buggs et al., 2014; Crawford et al., 2014; Garsmeur

et al., 2014; Barker et al., 2015), a better characterization of paleopolyploidies is an important next step in understanding the evolution of plant genomes.

Analyses of well assembled plant nuclear genomes have revealed the nature of some ancient polyploidies. Ploidal level has been inferred by searching for evidence of multiplied blocks of colinear genes in syntenic data from whole genome sequences or linkage mapping studies. For example, paleohexaploidy has been inferred in the ancestry of the core eudicots by identifying triplicated blocks of colinear genes (Tang et al., 2008). Similar evidence for paleohexaploidy has been inferred in the Solanaceae (Tomato Genome Consortium, 2012) and the tribe Brassiceae (Lagercrantz and Lydiate, 1996; Lysak et al., 2005; Tang et al., 2012). A recent high density linkage map of *Lactuca sativa* found evidence that the polyploidy in the ancestry of all Compositae (Barker et al., 2008) may be a paleohexaploidy (Truco et al., 2013). Truco et al. (2013) found that the *L. sativa* genome was triplicated when compared to the *Vitis vinifera* genome. However, their analyses could not phylogenetically place the set of genome duplications that generated the *Lactuca* paleohexaploidy. Using new data and phylogenomic methods, we more precisely placed this complex polyploidy in the history of the Compositae and closely related families.

Although most previous methods for identifying paleohexaploidy have used some form of syntenic or linkage data, a gene family approach with sufficient taxon sampling may also disentangle ancient hexaploidy and tetraploidy. Instead of identifying triplicated blocks of colinear genes, successive rounds of WGD should appear as phylogenetically nested duplications in nuclear gene family phylogenies. Higher ploidal level polyploidies, such as hexaploids, are formed by some combination of lower ploidal levels with genome doubling (Harlan and DeWet,

1975; Ramsey and Schemske, 1998). Hexaploids may be formed when a tetraploid crosses with a diploid. Triploids produced by these crosses may produce hexaploids when crossed themselves via unreduced 3*x* gametes. Alternatively, hexaploids may be formed from crosses between reduced (2*x*) and unreduced (4*x*) gametes from tetraploids with no diploids involved. Perhaps more rare are crosses of unreduced diploid (2*x*) and unreduced tetraploid (4*x*) gametes that may give rise to hexaploids. All of these pathways to hexaploids involve two, successive rounds of WGD that we may detect in the phylogenies of gene families from multiple species. Studying polyploidies that occurred tens of millions of years ago, it may be difficult to identify successive rounds of duplication if the WGDs happened very close in time or if the involved tetraploid or diploid lineages went extinct. However, if we have sampled extant representatives of the diploids or tetraploids involved in hexaploid formation, then we may place and infer the consecutive rounds of polyploidy giving rise to a paleohexaploid genome.

Using new data and algorithms, we re-evaluated our previous placement of the paleopolyploidy putatively shared by all extant Compositae (Barker et al., 2008). We sequenced new transcriptomes from the Barnadesieae (sister to all other Compositae), Calyceracae, Goodeniaceae, Menyanthaceae, and Campanulaceae. WGDs were inferred and placed in phylogenetic context using single species Ks plots of gene duplications (Barker et al., 2008, 2010) and our recently developed MAPS algorithm (Li et al., 2015). We also used MAPS to evaluate evidence of nested WGDs among our sampled taxa consistent with a paleotetraploidy followed by a paleohexaploidization.

MATERIALS AND METHODS

***Transcriptome sampling*—**We sequenced and assembled new transcriptomes for five species of Compositae and related families. Transcriptomes were sequenced for *Barnadesia spinosa* (Compositae), *Acicarpha spathulata* (Calyceraceae), *Scaevola aemula* (Goodeniaceae), *Nymphoides peltata* (Menyanthaceae), and *Platycodon grandiflorus* (Campanulaceae). *Barnadesia* was sequenced using Sanger technology, whereas the other transcriptomes were sequenced using 454 technology. RNA was extracted from leaves, flower buds, and mature flowers from each species. We isolated RNA using either Trizol reagent (Invitrogen, Carlsbad, CA), RNeasy kits (Qiagen, Valencia, CA), or a combination of the two methods. In the combined approach, the standard Trizol protocol was followed through the chloroform extraction step at which point 0.53X volumes of 100% ethanol was added to the aqueous phase and the entire RNA/ethanol mixture then applied to an RNeasy column and the Qiagen protocol followed thereafter. Approximately equal amounts of total RNA isolated from each tissue type were pooled prior to EST library preparation.

Two different methods were used to generate EST libraries. For Sanger sequencing, we prepared normalized libraries with the TRIMMER-DIRECT cDNA Normalization Kit (Evrogen, Moscow, Russia). cDNA samples from both the standard and normalized EST libraries were size-fractionated through agarose gels into three classes (0.5 - 1 kb, 1 – 2 kb, 2 – 3 kb) to reduce size biases during the subsequent cloning and sequencing steps. For 454 sequencing (454 Life Sciences, Branford, CT), we used modified oligo-dT primers during cDNA synthesis to reduce the length of mononucleotide runs associated with the polyA tail of mRNA. Mononucleotide runs reduce sequence quality and quantity due to excessive light production and crosstalk

between neighboring cells. For cDNA synthesis of the 454 libraries, we either used a broken chain short oligo-dT primer (Meyer et al., 2009) or two different modified oligo-dT primers: one to prime the polyA tail of mRNA during first strand cDNA synthesis and another to further break down the stretches of polyA sequence during second strand cDNA synthesis (Beldade et al., 2006; Lai, Zou, et al., 2012). We then normalized and amplified the cDNA using the TRIMMER-DIRECT cDNA Normalization Kit as above. After normalization, cDNA was sonicated to 500 to 800bp fragments and size-selected to remove small fragments. Then the fragmented ends were polished and ligated with adaptors. The optimal ligation products were selectively amplified and subjected to two rounds of size selection including gel electrophoresis and AMpure SPRI bead purification.

The *Barnadesia* Sanger EST library was sequenced using ABI 3730 machines (Life Technologies, Carlsbad, CA) at the Joint Genome Institute, Walnut Creek, CA. All other 454 libraries were sequenced on GS-FLX machines at Genome Quebec (http://www.genomequebec.com) using the standard Titanium Chemistry (http://454.com/). Raw reads for each newly sequenced transcriptome were cleaned and assembled into contigs. We used the SnoWhite cleaning pipeline (Dlugosch et al., 2013) to remove adapters, chimeras, and other possible contaminating sequences from raw reads. The cleaned reads were assembled using MIRA with the recommended settings for Sanger or 454 assemblies (Chevreux et al., 2004). Following MIRA, we finished the transcriptome assemblies with CAP3 (Huang and Madan, 1999) using a percent overlap of 94%. We also downloaded Sanger sequenced EST data from GenBank for *Artemisia annua (Nair et al., 2013)* and assembled with the same pipeline. Our previous transcriptome assemblies for *Helianthus annuus*, *Lactuca sativa*, and *Centaurea*

solstitialis (Barker et al., 2008; Lai, Kane, et al., 2012) were also used in this analysis. All raw and assembled data are available through the Compositae Genome Project Database (http://compgenomics.ucdavis.edu/compositae_database.php).

***DupPipe analyses of WGDs from transcriptomes of single species*—**For each transcriptome, we used our DupPipe pipeline to construct gene families and estimate the age of gene duplications (Barker et al., 2008, 2010). We translated DNA sequences and identified reading frames by comparing the Genewise alignment to the best hit protein from a collection of proteins from 25 plant genomes from Phytozome (Goodstein et al., 2012). For all DupPipe runs, we used protein-guided DNA alignments to align our nucleic acids while maintaining reading frame. We estimated synonymous divergence (Ks) using PAML with the F3X4 model (Yang, 2007) for each node in our gene family phylogenies. We identified peaks of gene duplication as evidence of ancient WGDs in histograms of the age distribution of gene duplications (Ks plots). We used a mixture model, EMMIX (McLachlan et al., 1999), to identify significant peaks and estimate their median Ks values.

***MAPS analyses of WGDs from transcriptomes of multiple species*—**To further infer and locate paleopolyploidy in our data sets, we used our recently developed gene tree sorting and counting algorithm, the Multi-tAxon Paleopolyploidy Search (MAPS) tool (Li et al., 2015). The MAPS algorithm uses a given species tree to filter collections of nuclear gene trees for subtrees consistent with relationships at each node in the species tree. Using this filtered set of subtrees, MAPS identifies and records nodes with a gene duplication shared by descendant taxa. To infer

and locate a potential whole genome duplication, we plotted the percentage of gene duplications shared by descendant taxa by node. A WGD will produce a large burst of shared duplications across taxa and gene trees. This burst of duplication will appear as an increase in the percentage of shared gene duplications in our MAPS analyses (Li et al., 2015). If the inferred WGDs are not too saturated, they will also appear in Ks plots for each species and corroborate the phylogenetic placement of WGDs. For our MAPS analysis, we used transcriptomes from five Compositae species and four outgroup species each representing closely related families of the Compositae. Our species tree for the MAPS analysis was based on previously published phylogenies (Kim et al., 2005; Panero and Funk, 2008; Funk et al., 2009; Tank and Donoghue, 2010; Soltis et al., 2011; Panero et al., 2014).

We circumscribed and constructed nuclear gene family phylogenies from multiple species for the MAPS analysis. We translated each transcriptome into amino acid sequences using the TransPipe pipeline (Barker et al., 2010). Using these translations, we performed reciprocal protein BLAST (blastp) searches among data sets for the MAPS analysis using an E-value of 10e-5 as a cutoff. We clustered gene families from these BLAST results using OrthoFinder under the default parameters (Emms and Kelly, 2015). Using a custom perl script, we filtered for gene families that contained at least one gene copy from each taxon and discarded the remaining OrthoFinder clusters. We used PASTA for automatic alignment and phylogeny reconstruction of gene families (Mirarab et al., 2015). For each gene family phylogeny, we ran PASTA until we reached three iterations without an improvement in likelihood score using a centroid breaking strategy. We constructed alignments using MAFFT (Katoh et al., 2002), employed Opal for mergers (Wheeler and Kececioglu, 2007), and RAxML for tree estimation

(Stamatakis, 2014). We used the best scoring PASTA tree for each multi-species nuclear gene family to infer and locate WGDs using MAPS.

RESULTS

Our newly sequenced and assembled transcriptomes for the Compositae and closely related families (Table 1) revealed a history of polyploidy. Peaks consistent with paleopolyploidy were observed in all Compositae and Calyceraceae species (Fig. 1). We previously identified WGD peaks ranging from Ks = 0.5–0.95 in the age distributions of *Centaurea*, *Helianthus*, and *Lactuca*, as well as an additional more recent WGD in the Heliantheae (Barker et al., 2008). The current analyses support these results and also find evidence for a WGD in the gene age distributions of *Artemisia* (Fig. 1A; median peak Ks = 0.86) and *Barnadesia* (Fig. 1B; median peak Ks = 0.56). Outside of the Compositae, we observed a similarly saturated WGD peak in the age distribution for *Acicarpha* from the Calyceraceae (Fig. 1C; median peak Ks = 0.56). The transcriptomes of the other outgroup taxa representing Goodeniaceae, Menyanthaceae, and Campanulaceae did not contain evidence of a WGD within a Ks < 2 window.

We used MAPS (Li et al., 2015) to test if the Compositae and Calyceraceae share an ancient WGD or experienced independent WGDs. Gene family clustering across the nine transcriptomes from species of Compositae and related families recovered 1,814 nuclear gene families with at least one copy from each sampled species. Among these gene families, MAPS revealed two bursts of shared gene duplication across the phylogeny (Fig. 2). One of these peaks of shared duplications was located in the MRCA of the Compositae and Calyceraceae. Forty

percent of the gene subtrees consistent with the species tree supported a gene duplication event at this node (Fig. 2 & 3). A burst of shared gene duplications at this node is consistent with the observed pattern of peaks from the single species Ks plots; inferred WGDs in *Acicarpha*, *Barnadesia*, and all other Compositae, but no apparent WGDs in the representatives of other families (Fig. 1).

We also observed a second burst of shared gene duplications among the core members of the Compositae (Fig. 2 & 3). More than 45% of the gene subtrees supported a shared duplication in the MRCA of *Centaurea*, *Lactuca*, *Artemesia*, and *Helianthus*. Although not easily distinguished in the Ks plots, this second round of WGD is consistent with linkage mapping data from *Lactuca* that indicated the Compositae WGD was actually a triplication (Truco et al., 2013). The two rounds of WGD observed in our MAPS analyses is consistent with a shared paleotetraploidization in the MRCA of the Compositae and Calyceraceae with a nested paleohexaploidization in the ancestry of the core Compositae.

DISCUSSION

Our analyses reveal a complex history of paleopolyploidy in the largest family of flowering plants. Single species (Ks plots) and multi-species phylogenomic analyses (MAPS) both place the previously identified Compositae WGD (Barker et al., 2008) in the ancestry of the Compositae and its sister family Calyceraceae. We did not find evidence of the WGD in any of the more closely related families of flowering plants, such as Goodeniaceae, Menyanthaceae, and Campanulaceae. However, we uncovered evidence for a second round of polyploidy among the core Compositae we analyzed, including representatives of *Helianthus*, *Lactuca*, *Artemisia*, and

Centaurea. Although not readily apparent in our Ks plots, our phylogenomic approach is able to disentangle the different duplication events in the nuclear gene family phylogenies. Notably, our Ks plots and MAPS analyses did not find evidence of a more recent WGD in *Artemisia* (Anthemideae) and support our previous placement of the putative Heliantheae WGD (Baldwin et al., 2002; Barker et al., 2008). Given the phylogenetic position of these inferred WGDs, our results indicate that most species of Compositae experienced at least two rounds of paleopolyploidy in the last ~50 million years.

The phylogenetically nested WGDs are consistent with our expectations for a paleohexaploidy in the ancestry of the Compositae. A high density linkage map of *Lactuca sativa* had previously found that the lettuce genome was triplicated (Truco et al., 2013). In nature, hexaploids are not formed from a single, instant triplication event, but rather consecutive genome duplications (Harlan and DeWet, 1975; Ramsey and Schemske, 1998). For a paleohexaploidy, the consecutive duplications appear as triplications in linkage map or syntenic data. However, it is difficult to distinguish these independent events in Ks plots because the duplications often occur too close in time to resolve, especially after many millions of years of evolution. If extant lineages survive from both duplication events, then it may be possible to resolve the two rounds of duplication in gene family phylogenies. That appears to be the case in the Compositae. Our analyses indicate that the Calyceraceae and at least tribe Barnadesieae descend from the first round of WGD, whereas the more derived Compositae in our analyses have a second polyploidy in their ancestry. Although we cannot resolve the exact mechanism of hexaploid formation with our present analyses, we can assign putative ploidal levels to these WGDs. The total evidence available from our present analyses and previous research (Barker et

al., 2008; Truco et al., 2013; Scaglione et al., 2016) indicates that the Calyceraceae and Barnadesieae are likely descendants of a paleotetraploid and the more derived Compositae tribes have a paleohexaploid ancestor. Previous analyses of the tribe Brassiceae (Lagercrantz and Lydiate, 1996; Lysak et al., 2005; Tang et al., 2012), the Solanaceae (Tomato Genome Consortium, 2012), and the eudicot triplication (Tang et al., 2008; Jiao et al., 2012; Vekemans et al., 2012) have used various gene family approaches to characterize these paleohexaploidies. There is no evidence in these cases for extant descendants of the paleotetraploid lineage. The Compositae and Calyceraceae represent a unique biological example with surviving lineages of both the paleotetraploid and paleohexaploid ancestors.

Although not precisely placed in the Compositae phylogeny because of limited sampling, we find that the second round of WGD may co-occur with many significant events in the evolution of the Compositae. The paleohexaploidization occurs in the same region of the phylogeny as two chloroplast DNA inversions (Jansen and Palmer, 1987; Kim et al., 2005). These inversions arose simultaneously, or nearly so, and are found among all sampled Compositae except *Barnadesia*. Intriguingly, our second WGD is co-located in the same region as these classic chloroplast inversions. It may be that the plastid inversions were introduced from a related species during an ancient hybridization event that produced the paleohexaploid ancestor. Two different plastid ancestors of extant Compositae may explain the rapid and simultaneous appearance of these plastid inversions in the phylogeny. Ancient hybridization and hexaploidization may have captured a plastid lineage that would otherwise have gone extinct. At present, we do not have sufficient resolution from our nuclear transcriptomes to assess if the paleohexaploid was an allo- or autohexaploid, but such diagnoses will be an interesting area of

future research. Plastid inversions and other rearrangements have been widely used to define major lineages of plants (Downie and Palmer, 1992; Doyle et al., 1992; Raubeson and Jansen, 1992; Graham and Olmstead, 2000; Schwarz et al., 2015). It is possible that ancient allopolyploidy in the Compositae and other clades may have captured these chloroplast genome variants. Combining our nuclear genome based placements of paleopolyploidy with abrupt chloroplast rearrangements as well as novel mitochondrial gene transfers (Adams et al., 2000, 2001, 2002; Adams and Palmer, 2003; Brandvain et al., 2007) may be a useful approach to recognize potential paleoallopolyploidies.

The inferred paleohexaploidy also co-occurs with the evolution of Compositae tribes outside of South America. Evidence suggests that the Calyceraceae and Composite likely originated and initially diversified in present-day South America (Panero and Funk, 2008; Funk et al., 2009; Barreda et al., 2010). The early branching lineages of Compositae, such as Barnadesieae and Mutisieae, are largely restricted to South America. Most of the family's diversity is found among the tribes that left South America and evolved on other continents (Panero and Funk, 2008). All of these "out of South America" tribes, which account for 96% of the family's diversity, are descended from the inferred paleohexaploid ancestor. In contrast, most of the South American lineages are descended from the paleotetraploid ancestor. It is unclear from our present analyses whether the genetics of the paleohexaploidy contributed to increased net diversification, facilitated intercontinental movement and ecological expansion, or if the association is simply a coincidence in a large phylogeny. Further sampling will permit more precise placement of these WGDs to better evaluate these polyploidies and diversification of the Compositae. The survival of paleotetraploid and paleohexaploid lineages represents a unique

opportunity to study the evolutionary outcomes and consequences of polyploidy of different ploidal levels.

An immediate observation from our results is that the Compositae WGD is not perfectly correlated with the origin or diversification of the Compositae. A recent analysis found that our previously identified Compositae WGD (Barker et al., 2008) was "perfectly associated" with an increase in net diversification rate (Tank et al., 2015). Given the information available at the time and the sampling in Tank et al. (2015), this is not an inaccurate statement. However, our new results suggest that the initial paleotetraploidy may have little association with increased net diversification as the Calyceraceae consists of only ~25 species and only ~4% of Compositae species are among the early diverging South American lineages (Panero and Funk, 2008; Funk et al., 2009). The inferred paleohexaploidy is correlated with a previously identified increase in net diversification rate that occurred just above the Barnadesioideae (Smith et al., 2011). This location in the Compositae phylogeny is a "hot spot" of activity: plastid inversions, global expansion of the family, and now a paleohexaploidy. Determining if and how these events are related and ultimately influenced the diversity of the Compositae will require more transcriptomic and genomic sampling of the early diverging lineages. Analyses of diverse Compositae genomes will permit us to test if increased divergent resolution of genes (Werth and Windham, 1991; Lynch et al., 2000; Scannell et al., 2006; McGrath et al., 2014; Muir and Hahn, 2015) or other mechanisms of post-polyploid genome evolution (Arrigo and Barker, 2012) contributed to the diversification of the Compositae. Ultimately, functional analyses connecting WGD duplicated genes with key morphological traits of the Compositae, such as Chapman et al. (2008), are needed to reveal the contribution of paleopolyploidy to plant diversity. The complex

WGDs in the Compositae and Calyceraceae provide a unique opportunity to explore these hypotheses.

ACKNOWLEDGEMENTS

The authors thank Tao Shi for assistance with a preliminary analysis of these data, and Tara Hall for assistance with submitting sequences to GenBank. We also thank Quentin Cronk of UBC for the *Barnadesia* sample and Cindy Sayre of VanDusen Botanical Garden for providing access to *Nymphoides* material. Hosting infrastructure and services were provided by the Biotechnology Computing Facility (BCF) at the University of Arizona. Funding was provided by the National Science Foundation Plant Genome Research Program to L.H.R. (DBI-0820451) and M.S.B. (NSF-IOS-1339156).

LITERATURE CITED

Adams, K.L., D.O. Daley, Y.-L. Qiu, J. Whelan, and J.D. Palmer. 2000. Repeated, recent and diverse transfers of a mitochondrial gene to the nucleus in flowering plants. *Nature* 408: 354–357.

Adams, K.L., and J.D. Palmer. 2003. Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. *Molecular Phylogenetics and Evolution* 29: 380–395.

Adams, K.L., Y.-L. Qiu, M. Stoutemyer, and J.D. Palmer. 2002. Punctuated evolution of mitochondrial gene content: high and variable rates of mitochondrial gene loss and transfer

to the nucleus during angiosperm evolution. *Proceedings of the National Academy of Sciences of the United States of America* 99: 9905–9912.

Adams, K.L., M. Rosenblueth, Y.L. Qiu, and J.D. Palmer. 2001. Multiple losses and transfers to the nucleus of two mitochondrial succinate dehydrogenase genes during angiosperm evolution. *Genetics* 158: 1289–1300.

Arrigo, N., and M.S. Barker. 2012. Rarely successful polyploids and their legacy in plant genomes. *Current Opinion in Plant Biology* 15: 140–146.

Baldwin, B.G., B.L. Wessa, and J.L. Panero. 2002. Nuclear rDNA Evidence for Major Lineages of Helenioid Heliantheae (Compositae). *Systematic Botany* 27: 161–198.

Barker, M.S., N. Arrigo, A.E. Baniaga, Z. Li, and D.A. Levin. 2015. On the relative abundance of autopolyploids and allopolyploids. *The New Phytologist*. doi: 10.1111/nph.13698.

Barker, M.S., K.M. Dlugosch, L. Dinh, R.S. Challa, N.C. Kane, M.G. King, and L.H. Rieseberg. 2010. EvoPipes.net: Bioinformatic Tools for Ecological and Evolutionary Genomics. *Evolutionary Bioinformatics Online* 6: 143–149.

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

Barker, M.S., H. Vogel, and M.E. Schranz. 2009. Paleopolyploidy in the Brassicales: analyses of

the *Cleome* transcriptome elucidate the history of genome duplications in *Arabidopsis* and other Brassicales. *Genome Biology and Evolution* 1: 391–399.

Barreda, V.D., L. Palazzesi, M.C. Tellería, L. Katinas, J.V. Crisci, K. Bremer, M.G. Passalia, et al. 2010. Eocene Patagonia fossils of the daisy family. *Science* 329: 1621.

Beldade, P., S. Rudd, J.D. Gruber, and A.D. Long. 2006. A wing expressed sequence tag resource for *Bicyclus anynana* butterflies, an evo-devo model. *BMC Genomics* 7: 130.

Brandvain, Y., M.S. Barker, and M.J. Wade. 2007. Gene co-inheritance and gene transfer. *Science* 315: 1685.

Buggs, R.J.A., J.F. Wendel, J.J. Doyle, D.E. Soltis, P.S. Soltis, and J.E. Coate. 2014. The legacy of diploid progenitors in allopolyploid gene expression patterns. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences* 369: 20130354.

Cai, J., X. Liu, K. Vanneste, S. Proost, W.-C. Tsai, K.-W. Liu, L.-J. Chen, et al. 2014. The genome sequence of the orchid *Phalaenopsis equestris*. *Nature Genetics* 47: 65-72.

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.

Chapman, M.A., J.H. Leebens-Mack, and J.M. Burke. 2008. Positive selection and expression divergence following gene duplication in the sunflower *CYCLOIDEA* gene family. *Molecular Biology and Evolution* 25: 1260–1273.

Cheng, S., E. van den Bergh, P. Zeng, X. Zhong, J. Xu, X. Liu, J. Hofberger, et al. 2013. The *Tarenaya hassleriana* genome provides insight into reproductive trait and genome evolution of crucifers. *The Plant Cell* 25: 2813–2830.

Chevreux, B., T. Pfisterer, B. Drescher, A.J. Driesel, W.E.G. Müller, T. Wetter, and S. Suhai. 2004. Using the miraEST Assembler for Reliable and Automated mRNA Transcript Assembly and SNP Detection in Sequenced ESTs. *Genome Research* 14: 1147–1159.

Crawford, D.J., J.J. Doyle, D.E. Soltis, P.S. Soltis, and J.F. Wendel. 2014. Contemporary and future studies in plant speciation, morphological/floral evolution and polyploidy: Honouring the scientific contributions of Leslie D. Gottlieb to plant evolutionary biology. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences* 369: 20130341.

Cui, L., P.K. Wall, J.H. Leebens-Mack, B.G. Lindsay, D.E. Soltis, J.J. Doyle, P.S. Soltis, et al. 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Research* 16: 738–749.

deWet, J.M.J. 1980. Origins of Polyploids. *In* W. H. Lewis [Ed.] Polyploidy, 3–15. Plenum Press, New York, USA.

Dlugosch, K.M., Z. Lai, A. Bonin, J. Hierro, and L.H. Rieseberg. 2013. Allele identification for transcriptome-based population genomics in the invasive plant *Centaurea solstitialis*. *G3* 3: 359–367.

Downie, S.R., and J.D. Palmer. 1992. Use of Chloroplast DNA Rearrangements in Reconstructing Plant Phylogeny. *In* P. S. Soltis, D. E. Solits, and J. J. Doyle [Eds.]

Molecular Systematics of Plants, 14–35. Springer US.

Doyle, J.J., J.I. Davis, R.J. Soreng, D. Garvin, and M.J. Anderson. 1992. Chloroplast DNA inversions and the origin of the grass family (Poaceae). *Proceedings of the National Academy of Sciences of the United States of America* 89: 7722–7726.

Doyle, J.J., L.E. Flagel, A.H. Paterson, R.A. Rapp, D.E. Soltis, P.S. Soltis, and J.F. Wendel. 2008. Evolutionary genetics of genome merger and doubling in plants. *Annual Review of Genetics* 42: 443–461.

Edger, P.P., H.M. Heidel-Fischer, M. Bekaert, J. Rota, G. Glöckner, A.E. Platts, D.G. Heckel, et al. 2015. The butterfly plant arms-race escalated by gene and genome duplications. *Proceedings of the National Academy of Sciences of the United States of America* 112: 8362–8366.

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

Estep, M.C., M.R. McKain, D. Vela Diaz, J. Zhong, J.G. Hodge, T.R. Hodkinson, D.J. Layton, et al. 2014. Allopolyploidy, diversification, and the Miocene grassland expansion. *Proceedings of the National Academy of Sciences of the United States of America* 111: 15149–15154.

Funk, V.A., T. Stuessy, and R. Bayer. 2009. Systematics, evolution, and biogeography of Compositae. International Association for Plant Taxonomy, Vienna, Austria.

Garsmeur, O., J.C. Schnable, A. Almeida, C. Jourda, A. D'Hont, and M. Freeling. 2014. Two evolutionarily distinct classes of paleopolyploidy. *Molecular Biology and Evolution* 31: 448–454.

Goodstein, D.M., S. Shu, R. Howson, R. Neupane, R.D. Hayes, J. Fazo, T. Mitros, et al. 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research* 40: D1178–86.

Graham, S.W., and R.G. Olmstead. 2000. Evolutionary significance of an unusual chloroplast DNA inversion found in two basal angiosperm lineages. *Current Genetics* 3: 7–183.

Harlan, J.R., and J.M.J. DeWet. 1975. On Ö. Winge and a prayer: the origins of polyploidy. *The Botanical Review* 41: 361–390.

Hegarty, M.J., and S.J. Hiscock. 2008. Genomic clues to the evolutionary success of polyploid plants. *Current Biology* 18: R435–R444.

Huang, X., and A. Madan. 1999. CAP3: A DNA sequence assembly program. *Genome Research* 9: 868–877.

Ibarra-Laclette, E., E. Lyons, G. Hernández-Guzmán, C.A. Pérez-Torres, L. Carretero-Paulet, T.-H. Chang, T. Lan, et al. 2013. Architecture and evolution of a minute plant genome. *Nature* 498: 94–98.

Jaillon, O., J.-M. Aury, B. Noel, A. Policriti, C. Clepet, A. Casagrande, N. Choisne, et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm

phyla. *Nature* 449: 463–467.

Jansen, R.K., and J.D. Palmer. 1987. A chloroplast DNA inversion marks an ancient evolutionary split in the sunflower family (Asteraceae). *Proceedings of the National Academy of Sciences of the United States of America* 84: 5818–5822.

Jiao, Y., J. Leebens-Mack, S. Ayyampalayam, J.E. Bowers, M.R. McKain, J. McNeal, M. Rolf, et al. 2012. A genome triplication associated with early diversification of the core eudicots. *Genome Biology* 13: R3.

Jiao, Y., J. Li, H. Tang, and A.H. Paterson. 2014. Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. *The Plant Cell* 26: 2792–2802.

Jiao, Y., N.J. Wickett, S. Ayyampalayam, A.S. Chanderbali, L. Landherr, P.E. Ralph, L.P. Tomsho, et al. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* 473: 97–100.

Katoh, K., K. Misawa, K.-I. Kuma, and T. Miyata. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* 30: 3059–3066.

Kihara, H., and T. Ono. 1926. Chromosomenzahlen und systematische Gruppierung der Rumex-Arten. *Zeitschrift für Zellforschung und Mikroskopische Anatomie* 4: 475–481.

Kim, K.-J., K.-S. Choi, and R.K. Jansen. 2005. Two Chloroplast DNA Inversions Originated Simultaneously During the Early Evolution of the Sunflower Family (Asteraceae).

Molecular Biology and Evolution 22: 1783–1792.

Lagercrantz, U., and D.J. Lydiate. 1996. Comparative genome mapping in *Brassica*. *Genetics* 144: 1903–1910.

Lai, Z., N.C. Kane, A. Kozik, K.A. Hodgins, K.M. Dlugosch, M.S. Barker, M. Matvienko, et al. 2012. Genomics of Compositae weeds: EST libraries, microarrays, and evidence of introgression. *American Journal of Botany* 99: 209–218.

Lai, Z., Y. Zou, N.C. Kane, J.-H. Choi, X. Wang, and L.H. Rieseberg. 2012. Preparation of normalized cDNA libraries for 454 Titanium transcriptome sequencing. *Methods in Molecular Biology* 888: 119–133.

Li, Z., A.E. Baniaga, E.B. Sessa, M. Scascitelli, S.W. Graham, L.H. Rieseberg, and M.S. Barker. 2015. Early genome duplications in conifers and other seed plants. *Science Advances* 1: e1501084.

Lynch, M., A.G. Force, and Editor: Joseph Travis. 2000. The origin of interspecific genomic incompatibility via gene duplication. *The American Naturalist* 156: 590–605.

Lysak, M.A., M.A. Koch, A. Pecinka, and I. Schubert. 2005. Chromosome triplication found across the tribe Brassiceae. *Genome Research* 15: 516–525.

Marques, I., S.A. Montgomery, M.S. Barker, T.D. Macfarlane, J.G. Conran, P. Catalán, L.H. Rieseberg, et al. 2015. Transcriptome-derived evidence supports recent polyploidization and a major phylogeographic division in *Trithuria submersa* (Hydatellaceae, Nymphaeales). *The*

New Phytologist. doi: 10.1111/nph.13755.

Mayrose, I., S.H. Zhan, C.J. Rothfels, K. Magnuson-Ford, M.S. Barker, L.H. Rieseberg, and S.P. Otto. 2011. Recently formed polyploid plants diversify at lower rates. *Science* 333: 1257.

McGrath, C.L., J.-F. Gout, P. Johri, T.G. Doak, and M. Lynch. 2014. Differential retention and divergent resolution of duplicate genes following whole-genome duplication. *Genome Research* 24: 1665–1675.

McLachlan, G.J., D. Peel, K.E. Basford, and P. Adams. 1999. The EMMIX software for the fitting of mixtures of normal and t-components. *Journal of Statistical Software* 4: 2.

Meyer, E., G. Aglyamova, S. Wang, J. Buchanan-Carter, D. Abrego, J. Colbourne, B. Willis, and M. Matz. 2009. Sequencing and *de novo* analysis of a coral larval transcriptome using 454 GSFlx. *BMC Genomics* 10: 219.

Ming, R., R. VanBuren, Y. Liu, M. Yang, Y. Han, L.-T. Li, Q. Zhang, et al. 2013. Genome of the long-living sacred lotus (*Nelumbo nucifera* Gaertn.). *Genome Biology* 14: R41.

Mirarab, S., N. Nguyen, and T. Warnow. 2015. PASTA: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. Journal of Computational Biology 22: 377-386.

Muir, C.D., and M.W. Hahn. 2015. The limited contribution of reciprocal gene loss to increased speciation rates following whole-genome duplication. *The American Naturalist* 185: 70–86.

Nair, P., A. Misra, A. Singh, A.K. Shukla, M.M. Gupta, A.K. Gupta, V. Gupta, et al. 2013. Differentially expressed genes during contrasting growth stages of *Artemisia annua* for

artemisinin content. *PloS One* 8: e60375.

Otto, S.P., and J. Whitton. 2000. Polyploid incidence and evolution. *Annual Review of Genetics* 34: 401–437.

Panero, J.L., S.E. Freire, L. Ariza Espinar, B.S. Crozier, G.E. Barboza, and J.J. Cantero. 2014. Resolution of deep nodes yields an improved backbone phylogeny and a new basal lineage to study early evolution of Asteraceae. *Molecular Phylogenetics and Evolution* 80: 43–53.

Panero, J.L., and V.A. Funk. 2008. The value of sampling anomalous taxa in phylogenetic studies: major clades of the Asteraceae revealed. *Molecular Phylogenetics and Evolution* 47: 757–782.

Ramsey, J., and D.W. Schemske. 1998. Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annual Review of Ecology and Systematics* 29: 467–501.

Raubeson, L.A., and R.K. Jansen. 1992. Chloroplast DNA evidence on the ancient evolutionary split in vascular land plants. *Science* 255: 1697–1699.

Rensing, S.A., D. Lang, A.D. Zimmer, A. Terry, A. Salamov, H. Shapiro, T. Nishiyama, et al. 2008. The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science* 319: 64–69.

Scaglione, D., S. Reyes-Chin-Wo, A. Acquadro, L. Froenicke, E. Portis, C. Beitel, M. Tirone, et al. 2016. The genome sequence of the outbreeding globe artichoke constructed *de novo* incorporating a phase-aware low-pass sequencing strategy of F1 progeny. *Scientific Reports*

6: 19427.

Scannell, D.R., K.P. Byrne, J.L. Gordon, S. Wong, and K.H. Wolfe. 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* 440: 341–345.

Schlueter, J.A., P. Dixon, C. Granger, D. Grant, L. Clark, J.J. Doyle, and R.C. Shoemaker. 2004. Mining EST databases to resolve evolutionary events in major crop species. *Genome* 47: 868–876.

Schmutz, J., S.B. Cannon, J. Schlueter, J. Ma, T. Mitros, W. Nelson, D.L. Hyten, et al. 2010. Genome sequence of the palaeopolyploid soybean. *Nature* 463: 178–183.

Schwarz, E.N., T.A. Ruhlman, J.S.M. Sabir, N.H. Hajrah, N.S. Alharbi, A.L. Al-Malki, C.D. Bailey, and R.K. Jansen. 2015. Plastid genome sequences of legumes reveal parallel inversions and multiple losses of rps16 in papilionoids. *Journal of Systematics and Evolution* 53: 458–468.

Shi, T., H. Huang, and M.S. Barker. 2010. Ancient genome duplications during the evolution of kiwifruit (*Actinidia*) and related Ericales. *Annals of Botany* 106: 497–504.

Smith, S.A., J.M. Beaulieu, A. Stamatakis, and M.J. Donoghue. 2011. Understanding angiosperm diversification using small and large phylogenetic trees. *American Journal of Botany* 98: 404–414.

Soltis, D.E., V.A. Albert, J. Leebens-Mack, C.D. Bell, A.H. Paterson, C. Zheng, D. Sankoff, et al. 2009. Polyploidy and angiosperm diversification. *American Journal of Botany* 96:

336–348.

Soltis, D.E., R.J.A. Buggs, J.J. Doyle, and P.S. Soltis. 2010. What we still don't know about polyploidy. *Taxon* 59: 1387–1403.

Soltis, D.E., S.A. Smith, N. Cellinese, K.J. Wurdack, D.C. Tank, S.F. Brockington, N.F. Refulio-Rodriguez, et al. 2011. Angiosperm phylogeny: 17 genes, 640 taxa. *American Journal of Botany* 98: 704–730.

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

Sveinsson, S., J. McDill, G.K.S. Wong, J. Li, X. Li, M.K. Deyholos, and Q.C.B. Cronk. 2014. Phylogenetic pinpointing of a paleopolyploidy event within the flax genus (*Linum*) using transcriptomics. *Annals of Botany* 113: 753–761.

Tang, H., X. Wang, J.E. Bowers, R. Ming, M. Alam, and A.H. Paterson. 2008. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Research* 18: 1944–1954.

Tang, H., M.R. Woodhouse, F. Cheng, J.C. Schnable, B.S. Pedersen, G. Conant, X. Wang, et al. 2012. Altered patterns of fractionation and exon deletions in *Brassica rapa* support a two-step model of paleohexaploidy. *Genetics* 190: 1563–1574.

Tank, D.C., and M.J. Donoghue. 2010. Phylogeny and phylogenetic nomenclature of the Campanulidae based on an expanded sample of genes and taxa. *Systematic Botany* 35:

425–441.

Tank, D.C., J.M. Eastman, M.W. Pennell, P.S. Soltis, D.E. Soltis, C.E. Hinchliff, J.W. Brown, et al. 2015. Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. *The New Phytologist*. doi: 10.1111/nph.13491.

Tomato Genome Consortium. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485: 635–641.

Truco, M.J., H. Ashrafi, A. Kozik, H. van Leeuwen, J. Bowers, S.R.C. Wo, K. Stoffel, et al. 2013. An ultra-high-density, transcript-based, genetic map of lettuce. *G3* 3: 617–631.

Vekemans, D., S. Proost, K. Vanneste, H. Coenen, T. Viaene, P. Ruelens, S. Maere, et al. 2012. Gamma paleohexaploidy in the stem lineage of core eudicots: significance for MADS-box gene and species diversification. *Molecular Biology and Evolution* 29: 3793–3806.

Velasco, R., A. Zharkikh, J. Affourtit, A. Dhingra, A. Cestaro, A. Kalyanaraman, P. Fontana, et al. 2010. The genome of the domesticated apple (*Malus* × *domestica* Borkh.). *Nature Genetics* 42: 833–839.

Wang, X., H. Wang, J. Wang, R. Sun, J. Wu, S. Liu, Y. Bai, et al. 2011. The genome of the mesopolyploid crop species *Brassica rapa*. *Nature Genetics* 43: 1035–1039.

Werth, C.R., and M.D. Windham. 1991. A model for divergent, allopatric speciation of polyploid pteridophytes resulting from silencing of duplicate-gene expression. *The American*

Naturalist 137: 515–526.

Wheeler, T.J., and J.D. Kececioglu. 2007. Multiple alignment by aligning alignments. *Bioinformatics* 23: i559–i568.

Wood, T.E., N. Takebayashi, M.S. Barker, I. Mayrose, P.B. Greenspoon, and L.H. Rieseberg. 2009. The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences of the United States of America* 106: 13875–13879.

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*. doi: 10.1093/molbev/msv081.

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* 24: 1586–1591.

Table 1. Summary statistics for the newly sequenced and assembled transcriptomes from species of Compositae and closely related families. All data available from the Compositae Genome Project Database (http://compgenomics.ucdavis.edu).

Species	Family	Genbank ID	Sequencing Technology	# of Unigenes	Total Assembly Length (mb)
Barnadesia spinosa	Compositae (Asteraceae)	SAMN00166828	Sanger	19,420	13.5
Acicarpha spathulata	Calyceraceae		454	57,264	44.9
Scaevola aemula	Goodeniaceae		454	56,070	27.1
Nymphoides peltata	Menyanthaceae		454	55,281	30.2
Platycodon grandiflorus	Campanulaceae		454	58,924	32.7

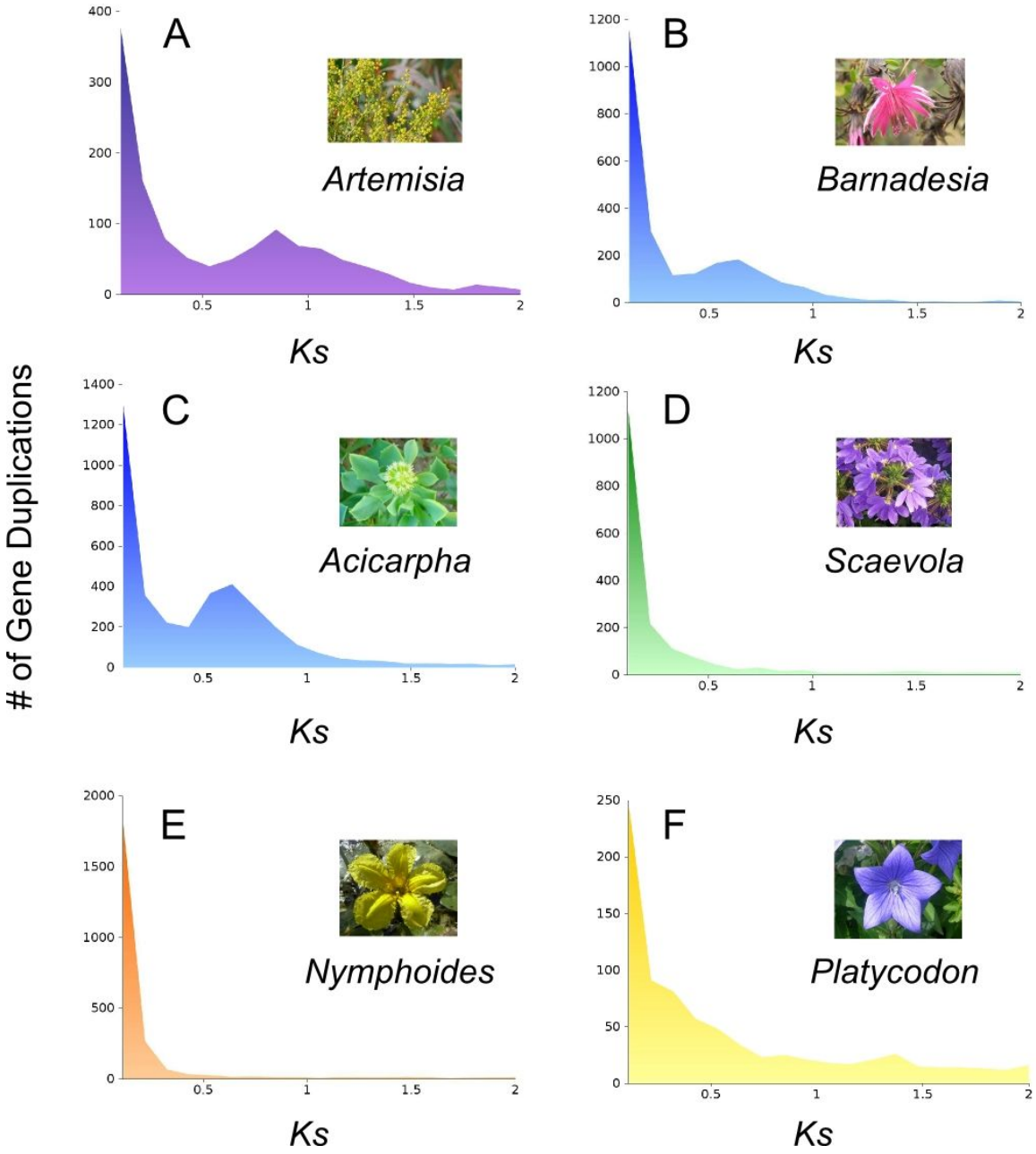

Figure 1. Histograms of gene age distributions for six species of Compositae and related families. Y-axis is number of gene duplications; x-axis is synonymous divergence (Ks) of the duplication. Peaks of gene duplication in A, B, and C correspond to WGDs. A. *Artemisia annua* (Compositae); B. *Barnadesia spinosa* (Compositae); C. *Acicarpha spathulata* (Calyceraceae); D. *Scaevola aemula* (Goodeniaceae); E. *Nymphoides peltata* (Menyanthaceae); F. *Platycodon*

grandiflorus (Campanulaceae). Image of *Barnadesia* provided by Quentin Cronk. Platycodon image from www.biolib.de and licensed for re-use under the Creative Commons Attribution-Share Alike License. Images A, C, D, and E from Wikipedia and licensed for re-use under the Creative Commons Attribution-Share Alike License. *Artemisia* image by Kristian Peters, *Acicarpha* image by Marcia Stefani, *Scaevola* image by Nemracc, and *Nymphoides* image by TeunSpaans.

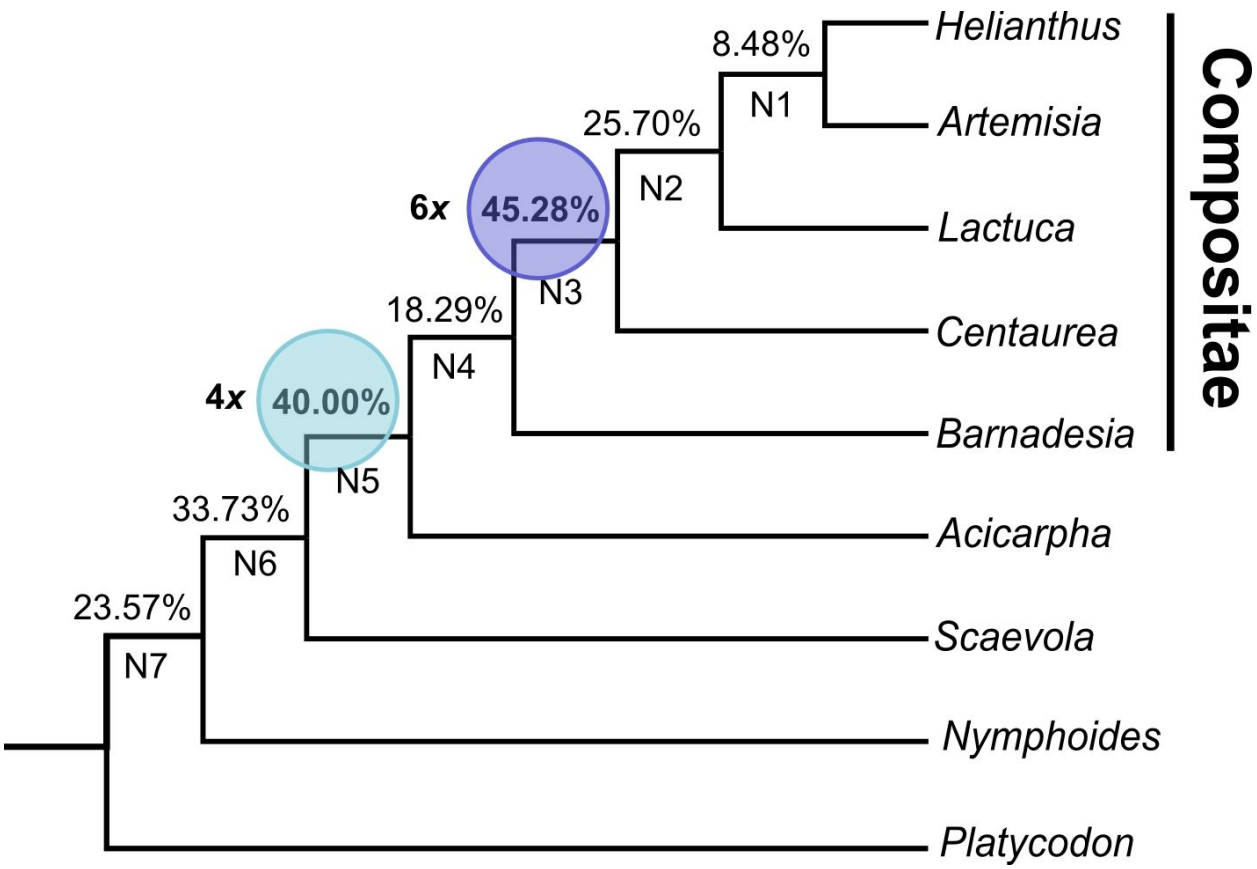

Figure 2. Species tree of Compositae and related families with % of gene duplications shared across 1,814 nuclear gene trees at each node, as inferred by MAPS. Colored circles highlight inferred WGDs at nodes N3 and N5. A paleotetraploidy is inferred in the ancestry of Compositae and Calyceraceae at node N5. A second round of WGD is inferred at node N3 and corresponds to putative paleohexaploidy in the ancestry of most Compositae.

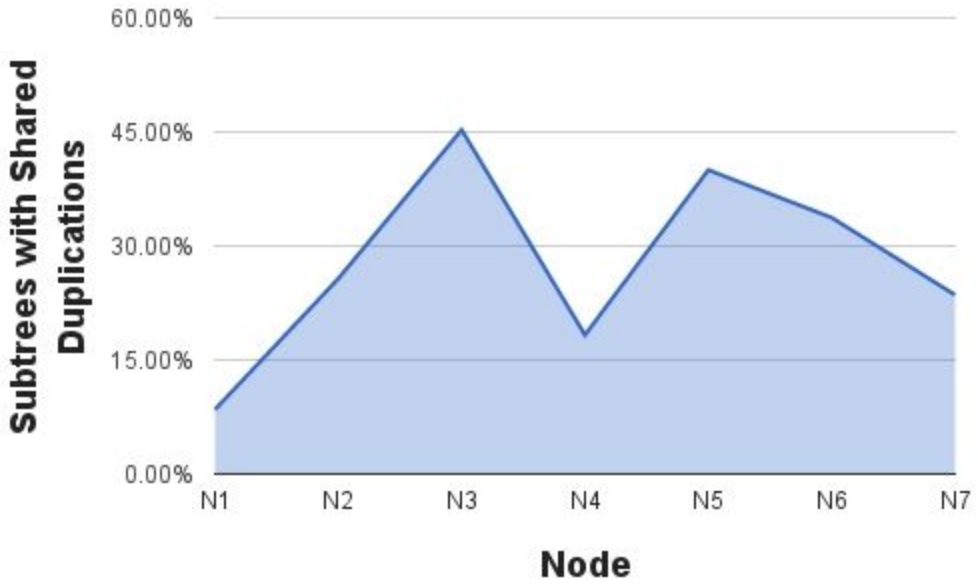

Figure 3. A histogram of the percentage of gene duplications shared across 1,814 nuclear gene families inferred by MAPS at each node in our analyzed species tree. The node numbers correspond to nodes in Figure 2. Two WGDs are inferred at the two peaks of shared gene duplication at nodes N3 and N5.