Article type : MS - Regular Manuscript

Improved transcriptome sampling pinpoints 26 ancient and more recent polyploidy events in Caryophyllales, including two allopolyploidy events

Ya Yang[1,4], Michael J. Moore[2], Samuel F. Brockington[3], Jessica Mikenas[2], Julia Olivieri[2], Joseph F. Walker[1] and Stephen A. Smith[1]

[1]Department of Ecology & Evolutionary Biology, University of Michigan, 830 North University Avenue, Ann Arbor, MI 48109-1048, USA; [2]Department of Biology, Oberlin College, 119 Woodland St, Oberlin, OH 44074-1097, USA; [3]Department of Plant Sciences, University of Cambridge, Cambridge, CB2 3EA, UK; [4]Current address: Department of Plant and Microbial Biology, University of Minnesota, Twin Cities. 1445 Gortner Avenue, St Paul, MN 55108, USA

Authors for correspondence:
Ya Yang
Tel: +1 612 625 6292
Email: yangya@umn.edu

Stephen A. Smith
Tel: +1 734 764 7923
Email: eebsmith@umich.edu

Received: *30 May 2017*
Accepted: *9 August 2017*

Summary

- Studies of the macroevolutionary legacy of polyploidy are limited by an incomplete sampling of these events across the tree of life. To better locate and understand these events, we need comprehensive taxonomic sampling as well as homology inference methods that accurately reconstruct the frequency and location of gene duplications.
- We assembled a dataset of transcriptomes and genomes from 169 species in Caryophyllales, of which 43 were newly generated for this study, representing one of the densest sampled genomic-scale datasets available. We carried out phylogenomic analyses using a modified phylome strategy to reconstruct the species tree. We mapped phylogenetic distribution of polyploidy events by both tree-based and distance-based methods, and explicitly tested scenarios for allopolyploidy.
- We identified 26 ancient and more recent polyploidy events distributed throughout Caryophyllales. Two of these events were inferred to be allopolyploidy.
- Through dense phylogenomic sampling, we show the propensity of polyploidy throughout the evolutionary history of Caryophyllales. We also provide a framework for utilizing transcriptome data to detect allopolyploidy, which is important as it may have different macro-evolutionary implications compared to autopolyploidy.

Key words: allopolyploidy, Caryophyllales, genome duplication, Ks plot, modified phylome, polyploidy.

Introduction

The prevalence and evolutionary consequences of polyploidy in plants have been discussed at length in the fields of macroevolution (Soltis *et al.*, 2015; Lohaus & Van de Peer, 2016). Polyploidy has been correlated with acceleration of speciation (Tank *et al.*, 2015; Smith *et al.*, 2017), surviving mass extinction (Fawcett *et al.*, 2009; Vanneste *et al.*, 2014a), evolutionary innovations (Vanneste *et al.*, 2014b; Edger *et al.*, 2015), and niche shift (Smith *et al.*, 2017). While there is little disagreement about the importance of polyploidy in angiosperm evolution, the frequency and phylogenetic locations of these events often remain unclear. Several limitations in methodology and sampling have limited our ability to accurately locate polyploidy events.

Until recently, most studies of polyploidy have employed either dating synonymous distances (Ks) among paralogous gene pairs (Vanneste *et al.*, 2013) or ancestral character reconstruction of chromosome counts (Mayrose *et al.*, 2010; Glick & Mayrose, 2014). While these have facilitated the discovery of many polyploidy events, both are indirect methods that have insufficient resolution and can be misleading (Kellogg, 2016). Ks plots between syntenic blocks from individual sequenced genomes have the advantage of being sensitive enough to detect ancient and nested polyploidy events (Jaillon *et al.*, 2007; Jiao *et al.*, 2011, 2012, 2014). However, this technique suffers from the typically sparse taxon sampling available in whole genome data. Distribution of polyploidy events inferred using Ks plots from genomic data, whether or not taking synteny into consideration (Fawcett *et al.*, 2009; Vanneste *et al.*, 2014a), await re-examination with more comprehensive taxon sampling. An alternative to Ks plots is the detection of polyploidy from chromosome counts. This method has the best signal for recent events and is most often restricted to the genus level or below (Wood *et al.*, 2009; Mayrose *et al.*, 2010, 2011).

Recent advances in transcriptome and genome sequencing offers the ability not only to measure Ks distances but also use gene tree topology to validate these. A combination of both approaches has allowed for the identification and placement of polyploidy events across the tree of life (Cannon *et al.*, 2015; Edger *et al.*, 2015; Li *et al.*, 2015; Marcet-Houben & Gabaldon, 2015; Yang *et al.*, 2015; Huang *et al.*, 2016; Xiang *et al.*, 2016). Despite this rapid increase in the number and precision of mapped polyploidy events, the sampling strategy for many of these studies was aimed at resolving deeper phylogenetic relationships. Testing hypotheses regarding the rich macroevolutionary legacy of polyploidy requires more extensive sampling of genomes and transcriptomes within a major plant clade. To date, only a few such data sets with sufficient sampling are available (Huang *et al.*, 2016; Xiang *et al.*, 2016). Furthermore, with a few exceptions (Kane *et al.*, 2009; Lai *et al.*, 2012; Estep *et al.*, 2014; Hodgins *et al.*, 2014), most of these studies have assumed autopolyploidy and have not explicitly tested for allopolyploidy. Despite the rich body of literature on gene expression, transposon dynamics, formation of novel phenotypes, and gene silencing and loss in recently formed allopolyploids (reviewed by Soltis & Soltis, 2016; Steige & Slotte, 2016), the long-term effects of allopolyploidy event remained poorly understood.

The plant order Caryophyllales offers an excellent opportunity to explore phylogenomic processes in plants. Caryophyllales forms a well-supported clade of *c*. 12500 species distributed among 39 families (Byng *et al.*, 2016; Thulin *et al.*, 2016), with an estimated crown age of *c*. 67–121 Ma (Bell *et al.*, 2010; Moore *et al.*, 2010; Smith *et al.*, 2017). Species of the Caryophyllales are found on every continent including Antarctica and in all terrestrial ecosystems as well as aquatic systems, occupying some of the most extreme environments on earth, including the coldest, hottest, driest, and most saline habitats inhabited by vascular plants. Familiar members of the group include cacti, living stones, a diverse array of carnivorous plants (e.g. the sundews, Venus flytrap, and tropical pitcher plants), and several important crop plants (e.g. beet, spinach, amaranth, and quinoa). Such extraordinary diversity makes Caryophyllales a prime system for investigating polyploidy vs diversification rate, character evolution, and niche shifts. Previous analyses using transcriptomes representing 67 species across Caryophyllales located 13 polyploidy events (Yang *et al.*, 2015). By generating 43 new transcriptomes we have expanded the previous sampling to include lineages with key evolutionary transitions, across a dataset that now includes 169 species of Caryophyllales.

The size of this dataset makes an all-by-all homology search impractical. Hence, we developed a 'modified phylome' strategy to build homolog and ortholog groups for species tree inference. In addition, we use an all-by-all approach to build lineage-specific homolog gene sets (Yang & Smith, 2014), and take advantage of recent developments in gene tree-based methods for mapping polyploidy events (Cannon *et al.*, 2015; Li *et al.*, 2015; Yang *et al.*, 2015). Our dense sampling allows us to take chromosome counts into consideration, and begin to explore allopolyploidy events. These improved methods for tree building and mapping of gene duplications, along with our improved taxon sampling, enable the most extensive exploration of polyploidy yet attempted in a major plan clade. The results reported here help establish the necessary foundation for further exploring the macroevolutionary consequences of polyploidy (for example, Smith *et al.*, 2017).

Materials and Methods

Taxon sampling, laboratory procedure, and sequence processing

We included 178 ingroup datasets (175 transcriptomes, three genomes; Supporting Information Table S1) representing 169 species in 27 out of the 39 Caryophyllales families (Byng *et al.*, 2016; Thulin *et al.*, 2016). Among these, 43 transcriptomes were newly generated for this study (Table S2). In addition, 40 outgroup genomes across angiosperms were used for rooting gene trees (Table S1). Tissue collection, RNA isolation, library preparation, sequencing, assembly, and translation followed previously published protocols (Brockington *et al.*, 2015; Yang *et al.*, 2017) with minor modifications (Tables S1, S2).

Caryophyllales homology and orthology inference from peptide sequences using a 'modified phylome' strategy

We employed a modified phylome strategy for reconstructing orthogroups (Fig. S1). An 'orthogroup' includes the complete set of genes in a lineage from a single copy in their common ancestor. Each node in an orthogroup tree can represent either a speciation event or a gene duplication event. An orthogroup differs from a homolog group in that the former is inferred from the latter by extracting rooted ingroup lineages separated by outgroups. The modified phylome procedure consisted of two major steps. First, 'backbone homolog groups' were constructed using peptide sequences from three Caryophyllales and 40 outgroup genomes. Second, peptides from transcriptomes were sorted to each backbone homolog. This two-step procedure allowed us to avoid the computationally intensive all-by-all homology search for constructing orthogroups.

To construct the backbone homolog groups, we started from the proteome of the best annotated genome, sugar beet (Fig. S1; Dohm *et al.*, 2014; http://bvseq.molgen.mpg.de/ v1.2, accessed June 25, 2015). Sequences from each beet locus were used to search against a database that consisted of combined proteomes from all 43 genomes using SWIPE v2.0.11 (Rognes, 2011) with an *E*-value cutoff of 0.01. The top 100 hits with bit scores higher than 50, and bit scores of at least 20% of the self-hit were retained and aligned using MAFFT v7.215 (Katoh & Standley, 2013), with '--genafpair --maxiterate 1000'. The alignments were trimmed using Phyutility v2.2.6 (Smith & Dunn, 2008) with '-clean 0.1', and trees were constructed using RAxML v8.1.5 (Stamatakis,

2014) with the model PROTCATWAG. After visual inspection of *c*. 10 resulting trees to evaluate outliers, terminal branches that were longer than two (absolute cutoff) or longer than one and more than 10 times as long as its sister (relative cutoff) were trimmed. Internal branches longer than one were separated (Yang & Smith, 2014). We retained trees that contained the original beet bait locus, combining those groups that shared beet locus IDs (i.e. had gene duplication within Caryophyllales).

The resulting backbone homolog groups constructed from the 43 genomes were then used to place the remaining 175 Caryophyllales transcriptomes (Fig. S1). First, peptide sequences from each of the Caryophyllales transcriptomes were reduced using CD-HIT v4.6 (-c 0.99 -n 5; Fu *et al.*, 2012). The resulting sequences were then used in SWIPE analyses comparing the sequences to the beet proteome to identify matching backbone homolog groups. A new tree representing each expanded homolog group, with both genome and transcriptome data, was estimated using the same alignment and phylogenetic reconstruction settings as for the backbone homolog tree. To reduce isoforms in transcriptome datasets, monophyletic and paraphyletic tips that belonged to the same taxon were removed, leaving only the tip with the highest number of characters in the trimmed alignment (Yang & Smith, 2014). Spurious tips and long internal branches were cut using the same settings as for the backbone tree. For homolog groups with >1000 and <5000 sequences, alignments were constructed using PASTA v1.6.3 (Mirarab *et al.*, 2014) with default settings, were trimmed by Phyutility with '-clean 0.01', and phylogenetic trees were estimated using FastTree v2.1.8 (Price *et al.*, 2010) with the model 'WAG'. An initial internal branch length cutoff of 2 was used after reducing tips and trimming spurious tips with the same cutoffs as for the backbone trees. A second round of alignment and refining was carried out for these larger homolog groups. Homolog groups larger than 5000 were ignored.

After obtaining final homologs using the modified phylome approach, we carried out orthology inference following the 'rooted ingroup' method in Yang & Smith (2014). Briefly, for each Caryophyllales orthogroup extracted from a final homolog, we walked from the root towards the tip. When two sister nodes share one or more taxa, the side with a smaller number of taxa was separated and both subtrees were taken into account in the next round until all subtrees contained only one sequence per taxon. For each resulting tree with at least 160 taxa, sequences were pooled, re-aligned using PRANK v140110 (Löytynoja & Goldman, 2010) with default settings, trimmed

with Phyutility with '-clean 0.3', and a new ortholog tree estimated using RAxML with 'PROTCATAUTO'. A set of more stringent cutoffs was used to produce the final ortholog trees: absolute tips cutoff of 0.6, relative tip cutoff of 0.3, and an internal branch cutoff of 0.4. Aligned sequences were pooled according to remaining tips, trimmed (Phyutility with '-clean 0.3'), and remaining alignments with at least 150 characters and 160 taxa were used for species tree inference.

All-by-all homology search and orthology inference in each of five Caryophyllales subclades from coding sequences

Uncertainty in alignment and tree inference increases with dataset size. Given the absence of polyploidy events along the backbone of Caryophyllales (Smith *et al.*, 2015; Yang *et al.*, 2015), we divided Caryophyllales into five subclades according to previous phylogenetic analysis (Yang *et al.*, 2015): PHYT: Aizoaceae+the 'Phytolaccoid clade' that consists of Nyctaginaceae, Phytolaccaceae s.l. (i.e. including *Agdestis*), Petiveriaceae, and Sarcobataceae (Yang *et al.*, 2015), with *Stegnosperma halimifolium* (Stegnospermataceae) and the three Caryophyllales genomes *Beta vulgaris* (beet, Chenopodiaceae; Dohm *et al.*, 2014), *Spinacia oleracea* (spinach, Chenopodiaceae; Dohm *et al.*, 2014), and *Dianthus caryophyllus* (carnation, Caryophyllaceae; Yagi *et al.*, 2014) as outgroups; PORT, the 'Portullugo clade' that consists of Molluginaceae+Portulacineae (Edwards & Ogburn, 2012) with the three Caryophyllales genomes as outgroups; AMAR, Amaranthaceae+Chenopodiaceae, with carnation and *Phaulothamnus spinescens* (Achatocarpaceae) as outgroups; CARY, Caryophyllaceae, with spinach, beet and *P. spinescens* (Achatocarpaceae) as outgroups; and NCORE, the clade that is sister to the rest of Caryophyllales, with all three Caryophyllales genomes plus *Microtea debilis*, *Physena madagascariensis* and *Simmondsia chinensis* as outgroups.

An all-by-all approach was used for homology inference in each subclade following Yang & Smith (2014) with minor modifications (Methods S1). The final alignments from homolog trees with no taxon duplication (i.e. one-to-one orthologs), no more than one missing taxon (except requiring full taxon occupancy for CARY and PHYT), and average bootstrap value of at least 80 were trimmed with Phyutility '-clean 0.5'. Trimmed alignment with at least 300 columns were the final orthologs.

Species tree inference

We used two alternative approaches for constructing species trees for both the entire Caryophyllales using peptides ('modified phylome dataset') and each of the five subclades using coding sequences (CDS) ('the subclade dataset'). First, a supermatrix was constructed by concatenating trimmed ortholog alignments. A maximum likelihood tree was estimated from the supermatrix using RAxML, partitioning each locus, with the model set to PROTCATAUTO for peptides and GTRCAT for coding sequences for each individual partition. Node support was evaluated by the internode certainty all (ICA) scores (Salichos *et al.*, 2014) calculated in RAxML using final ortholog trees as input. Probabilistic correction was used to take incomplete taxon occupancy into consideration (Kobert *et al.*, 2016; Stamatakis, 2016). As implementation of ICA score calculation was updated in more recent releases of RAxML, we used RAxML v. 8.2.9 for calculating ICA scores.

In addition to the concatenated analyses, we also searched for the maximum quartet support species tree (MQSST) using ASTRAL-II v. 4.10.12 (Mirarab *et al.*, 2014; Mirarab & Warnow, 2015) starting from maximum likelihood trees estimated from individual orthologs. Tree uncertainty was evaluated by using 100 multi-locus bootstrap replicates (Seo *et al.*, 2005; Seo, 2008; Mirarab *et al.*, 2014), starting from 200 fast bootstrap trees for each final ortholog calculated in RAxML.

Mapping polyploidy events based on subclade orthogroup tree topology

To map polyploidy events in each subclade, we extracted orthogroups from each subclade homolog tree, requiring no more than two missing ingroup taxa. When two or more taxa overlapped between the two daughter clades, a gene duplication event was recorded to the most recent common ancestor (MRCA) on the subclade species tree (Yang *et al.*, 2015). In this procedure, each node on a species tree can be counted at most once per orthogroup to avoid nested gene duplications inflating the number of duplications scored. Two alternative filters were applied for comparison. The first filter required an average bootstrap percentage of each orthogroup to be at least 50. Alternatively, we also tested a local topology filter that only mapped a gene duplication

event when the sister clade of the gene duplication node in the orthogroup contained a subset of the taxa in the corresponding sister clade in the species tree (Cannon *et al.*, 2015; Li *et al.*, 2015).

Distribution of synonymous distance among gene pairs (Ks plots)

For each of the ingroup Caryophyllales datasets, a Ks plot of within-taxon paralog pairs was created following the same procedure as Yang *et al*. (2015) based on BLASTP hits. Similarly, we carried out a second Ks analysis based on BLASTN between CDS without first reducing highly similar sequences to maximize detection of more recent polyploidy events. In cases where tree-based mapping was ambiguous, or comparison of within- vs between-species Ks peaks could help inform allopolyploidy, we also calculated Ks distribution of between-species reciprocal best BLASTN hit pairs. Ks values <0.01 were excluded to avoid isoforms from *de novo* assembled transcriptomes.

Chromosome counts

Chromosome counts were obtained from the Chromosome Counts Database (ccdb.tau.ac.il accessed 5 October 2015). When counts in this database were unavailable or inconsistent, counts were obtained from the Jepson eFlora (ucjeps.berkeley.edu/eflora/ accessed 5 October 2015) and Flora of North America (www.efloras.org/ accessed 5 October 2015).

A total evidence approach for mapping polyploidy events

We considered six scenarios for mapping polyploidy events including taking orthogroup tree topology, within-taxon Ks plots, and chromosome counts into consideration (Fig. 1). When polyploidy events occurred without subsequent speciation (or for which only one taxon is represented in our sampling; Fig. 1a–c), only a single within-taxon Ks plot would show a peak. In these instances, because we required at least two overlapping taxa between sister clades in the orthogroup tree to record a gene duplication event, no gene duplication was recorded from topology-based mapping. The polyploidy event was therefore mapped to the terminal branch with the Ks peak. However, when a polyploidy event was followed by lineage diversification, we used information from both Ks peaks and the orthogroup tree topologies to map duplication events (Fig. 1d–f). If we saw an excess of duplication events along the same lineage in which all taxa share the

same within-taxon Ks peak, then this was inferred as autopolyploidy (Fig. 1d). However, if an excess of duplication events was found on a lineage ancestral to the lineage in which all taxa share the same within-taxon Ks peak, this was inferred to be an allopolyploidy event (Fig. 1e,f). We indicated the node with low support in Fig. 1 to highlight that allopolyploidy can lead to nodes with low support when both parental lineages are present or nodes that are well supported when one of the parental lineages is missing. We did not consider polyploidy events that were supported by chromosome count alone, as these can be within-population variation (Caperta *et al.*, 2016), and an increase in number can represent chromosome fission instead of duplication (Fishman *et al.*, 2014; Chester *et al.*, 2015).

Results

Data availability

Raw reads for newly generated transcriptomes were deposited in the NCBI Sequence Read Archive (BioProject: PRJNA388222; Table S2). Assembled sequences, alignments, and trees were deposited in Dryad (doi:10.5061/dryad.st3gt). Scripts used were also archived in Dryad, with notes and updates for modified phylomes available from https://bitbucket.org/yangya/genome_walking_2016 and those for building lineage-specific homologs and mapping polyploidy events available from https://bitbucket.org/blackrim/clustering.

RAxML and ASTRAL recovered nearly identical species tree topologies

Both RAxML and ASTRAL analyses recovered identical topologies for most branches (Figs 2–6, S2–S4). We consider branches with an ICA score higher than 0.5 as strongly supported, as ICA scores lower than 0.5 suggests that the dominant bipartition is present in <80% of ortholog trees (Salichos *et al.*, 2014). As multi-locus bootstrap support percentages increase with the number of loci (Seo, 2008) and given that each of our final ortholog set contained more than a hundred loci (Table 1), we consider multi-locus bootstrap values <100 as low support. Using this set of criteria, most branches from subclade datasets (Figs 2–6, S2) and the majority of the branches from modified phylomes (Figs S3, S4) were well-supported.

We recovered between 152 to 736 one-to-one orthologs and 0.2–1.1 million trimmed CDS columns from each of the five subclades. The concatenated supermatrices had gene occupancies of 98–100% and character occupancies of 87–93% (Table 1). Four clades showed different relationships between RAxML and ASTRAL, with little support for either alternative relationships (Figs 2–6 marked with '*', S2): *Cyphomeris gypsophiloides* (Nyctaginaceae, PHYT) was sister to *Allionia* in the RAxML tree (ICA = -0.01) but was sister to the clade *Nyctaginia+Anulocaulis+Boerhavia* in the ASTRAL tree (bootstrap = 95); species in *Leuenbergeria* (Cactaceae, PORT) were monophyletic in the RAxML tree (ICA = -0.09) but were paraphyletic to the rest of Cactaceae in the ASTRAL tree (bootstrap = 69); *Tidestromia lanuginosa* (Amaranthaceae, AMAR) was sister to the clade of *Froelichia+Guilleminea+Gossypianthus+Blutaparon+Alternanthera* in the RAxML tree (ICA = -0.00), but was sister to *Alternanthera* in the ASTRAL tree (bootstrap = 50); and *Saponaria officinalis* (Caryophyllaceae, CARY) was sister to *Gypsophila+Dianthus+Velezia* in the RAxML tree (ICA = -0.04), but was sister to *Dianthus+Velezia* in the ASTRAL tree (bootstrap = 63).

Among the 15045 homolog groups we obtained using the modified phylome approach, 15 had >5000 sequences and were ignored, while the rest were used for subsequent orthology inference. The final concatenated matrix consisted of 624 loci and 215,669 amino acids, with a final gene occupancy of 92.6% and character occupancy of 80.1% (Table 1). The modified phylome approach recovered identical species tree topologies except for one branch that had little support from either analysis (Figs S3, S4): *Leuenbergeria* (Cactaceae) was monophyletic in the RAxML tree (ICA = 0.18) but was polyphyletic in the ASTRAL tree (bootstrap = 28). The modified phylome approach recovered an identical species tree topology compared to that recovered by the subclade analysis. When subclade trees had different topologies between RAxML and ASTRAL, the modified phylome tree agreed with the subclade ASTRAL results in the placement of *Cyphomeris gypsophiloides* (ICA = 0.23 and bootstrap = 97), whereas the position of *Leuenbergeria* was recovered in the same incongruent positions as recovered by RAxML and ASTRAL in the subclade tree analyses. The modified phylome approach recovered both *Tidestromia lanuginosa* (0.19/81) and *Saponaria officinalis* (0.38/98) in the same positions as found in the RAxML results for the subclade trees.

Twenty-six polyploidy events were mapped

Twenty-six polyploidy events were inferred by using a total evidence approach of orthogroup tree topology, shared Ks peaks, and chromosome counts (Figs 2–6). Overall the two orthogroup tree filtering strategies, by bootstrap percentage or local tree topology, produced almost identical results for frequency of gene duplication (Figs 2–6, S5). The frequency of gene duplications was strongly associated with the inferred polyploidy events (Figs 2–6). In our Ks analysis, we only considered Ks peaks that were similar in height or taller than the peak, *c*. Ks = 2, that corresponds to the early eudicot paleohexaploidy that predates the origin of Caryophyllales (Dohm *et al.*, 2012; Jiao *et al.*, 2012; Yang *et al.*, 2015).

Two polyploidy events in the PHYT clade were supported by both homolog tree topology and shared Ks peaks (Figs 2, S5a, S6a). The frequency of orthogroups showing evidence of gene duplication were 52% filtered by bootstrap percentage and 45% filtered by tree topology for PHYT1, and 33% and 31% respectively for PHYT2. Both were significantly higher percentages compared to remaining branches (Fig. 2).

At least four polyploidy events were recovered in the PORT clade. PORT1 was mapped to both the MRCA of Portulacineae (19%/17%) and its parent node (24%/22%) from gene duplications (Figs 3, S5b). However, Molluginaceae did not share the Ks peak that was present in all members of Portulacineae (Fig. S6b). Within-species paralogs in Portulacineae (represented by the PORT1 Ks peak in *Talinum* sp. at Ks = 0.64; Fig. 3, lower left) coalesced at lower Ks values compared to the Ks peak at 0.76 between *Talinum* sp. and *Mollugo pentaphylla* (Molluginaceae). However, similar comparison of *Portulaca pilosa* (Portulacineae) vs *Mollugo pentaphylla* showed overlapping Ks peaks (Ks = 0.9; Fig. 3, lower right), likely due to faster molecular rate in *Portulaca* compared to *Talinum*. Therefore, phylogenetic uncertainty likely at least partly contributed to the ambiguity in mapping. Both PORT2 and four were recovered by taxon-specific Ks peaks, and both had relatively high chromosome counts compared to close relatives (Figs 3, S6b). PORT3 was supported by shared Ks peaks and gene duplications in orthogroup trees (21%/18%), whereas chromosome counts were uninformative.

At least three polyploidy events were recovered in the AMAR clade (Figs 4, S5c, S6c). AMAR1 was detected by an elevated percentage of gene duplications mapped to the branch

uniting *Alternanthera+Gossypianthus+Blutaparon+Froelichia+Aerva* (37%/35%). However, species of *Aerva* lacked the AMAR1 Ks peak shared by *Alternanthera+Gossypianthus+Blutaparon+Froelichia* at 0.4–0.65. Further examination of the between-species Ks peak of *Aerva javanica* vs *Tidestromia lanuginose* (Ks = 0.46) shows that it was more ancient than the within-species Ks peak AMAR2 at 0.23 in *Aerva javanica* (Fig. 4, lower left), suggesting that paralogs in *Aerva javanica* coalesced more recently than coalescing with taxa outside of *Aerva*, and AMAR1 and 2 were two distinct polyploidy events. The Ks peak within *Aerva javanica* (AMAR2, Ks = 0.23) overlapped with the between-species Ks peak of *A. javanica* vs *A. lanata* (Ks = 0.24; Fig. 4, lower right). Given that *A. javanica* had faster molecular substitution rate than *A. lanata* according to their relative branch lengths (Fig. 4), paralogous copies within *A. javanica* likely coalesced along the branch leading to *A. javanica* before coalescing with *A. lanata*. The lack of the AMAR2 peak in *A. lanata* as well as chromosome counts (2*n* = 32 for *A. javanica* vs 2*n* = 16 for *A. lanata*) further supported the location of AMAR2 along the terminal branch leading to *A. javanica.* Based on the lack of the AMAR1 peak in both *Aerva* species, we inferred that AMAR1 was an allopolyploidy event, with one parental lineage closely related to *Aerva* and the other parental lineage missing, consistent with the scenario in Fig. 1(f).

At least seven polyploidy events were recovered in the CARY clade (Figs 5, S5d, S6d). CARY1 was detected through an elevated percentage of gene duplication in two adjacent nodes (21%/20% on the node that included *Spergularia media*, and 23%/20% in the node excluded *S. media*; Figs 5, S5d). However, *S. media* did not share the CARY1 Ks peak (Fig. S6d). The reciprocal best hits Ks peak between *Spergularia media* and *Silene latifolia* indicated that paralogs derived from CARY1 coalesced within *Silene latifolia* at similar Ks values compared to coalescing with *Spergularia media* (Fig. 5), suggesting that phylogenetic uncertainty at least partly contributed to the fact that CARY1 mapped to two adjacent nodes.

Nested in CARY1, five taxa showed Ks peaks (Figs 5, S6d). Among them, within-species Ks peak CARY2 was observed only in *Cerastium fontanum* (2*n* = 72) but missing from its sister *C. arvense* (2*n* = 36). *Honckenya peploides* had a within-species Ks peak *c*. 0.06 (CARY3), its sister *Schiedea membranacea* had a Ks peak at 0.22 (CARY4), whereas their reciprocal best hit Ks peak was at 0.08, suggesting that CARY3 was restricted to the terminal branch leading to *H. peploides*.

The observation that the within-species Ks peak in *S. membranacea* (CARY4) was older than its split with *H. peploides*, but that CARY4 was not shared with *H. peploides*, suggested an allopolyploid origin of *S. membranacea* with one parental lineage closely related to *H. peploides* while the other parental lineage missing from our taxon sampling. Pairwise comparison among *S. membranacea*, *H. peploides*, and *Scleranthus polycarpos* showed that *S. polycarpos* had a polyploidy event CARY5 (Ks = 0.04) that is more recent than its split with *Schiedea* or *Honckenya* and is therefore restricted to *S. polycarpos*. CARY6, also nested in CARY1, was mapped to the terminal branch leading to *Colobanthus quitensis*. Although *Colobanthus* was weakly supported to be sister to the clade consisted of *Honckenya*+*Schiedea*+*Scleranthus* (ICA = 0.13, bootstrap = 100), the peak at Ks = 0.15 was more recent than the *Honckenya*/*Scleranthus* (Ks = 0.24) or the *Schiedea*/*Scleranthus* (Ks = 0.27) split and is therefore inferred to be independent of CARY 3, 4, or 5. In addition to CARY2–6 that were nested in CARY1, one additional polyploidy event CARY7 independent of CARY1 was mapped to the MRCA of *Drymaria cordata* and *D. subumbellata* by a shared Ks peak. Both species had high chromosome counts relative to their sister lineages.

At least six polyploidy events were inferred in the NCORE clade (Figs 6, S5e, S6e). Both gene duplications (43%/34%) and shared Ks peaks supported a polyploidy event at the base of Polygonaceae (NCORE1). NCORE2–5 were supported by Ks peaks, and NCORE6 was supported by both Ks peaks and chromosome counts. We inferred NCORE5 (base of Droseraceae) and NCORE6 (branch leading to *Nepenthes alata*, Nepenthaceae) as two separate polyploidy events on sister branches given that very low frequencies of gene duplication events were mapped to the MRCA of Droseraceae+Nepenthaceae (0.9%/1.7%), compared to the MRCA of Droseraceae (2.8%/3.0%) and Nepenthaceae (16%/15%).

In addition to the polyploidy events detected from each of the five subclades, three of the four taxa along the grade paraphyletic to PHYT+PORT+AMAR+CARY also each had a peak at Ks lower than 1: *S. halimifolium* (Fig. S6a), *P. madagascariensis* (Fig. S6e), and *S. chinensis* (Fig. S6e). No polyploidy event has been inferred along the Caryophyllales backbone leading to beet (Chenopodiaceae) from genome analysis (Dohm *et al.*, 2012; Dohm *et al.*, 2014), indicating Ks peaks mapped to this grade likely represent lineage-specific polyploidy events. Also, the relatively high chromosome count of *S. chinensis* (2*n* = 52) compared to *M. debilis* (2*n* = 18), the only taxon in this grade that did not experience a polyploidy event, further supports the lineage specific nature

of the polyploidy events along this grade. In addition, *P. spinescens* (Achatocarpaceae; sister to AMAR) also likely had its lineage specific polyploidy event, as it had a Ks peak that was not shared with its sister clade (Figs 4, S6c).

Since we excluded Ks values <0.01 when plotting to avoid isoforms in *de novo* assembled transcriptome data resulted in very high Ks counts, occasionally there were apparent peaks at Ks < 0.2. We have plotted all Ks values on a different scale to zoom in on these Ks peaks (shown in Fig. S6 only when Ks peak < 0.2 was confirmed).

Discussion

Our analyses add to a growing body of literature that suggests that polyploidy events are much more prevalent than previously thought (Cannon *et al.*, 2015; Edger *et al.*, 2015; Li *et al.*, 2015; Yang *et al.*, 2015; Huang *et al.*, 2016; Xiang *et al.*, 2016; Mandakova *et al.*, 2017). The dataset presented here uniquely contributes to this question by greatly improving taxon sampling of transcriptomes in a major plant clade (169 species in Caryophyllales) whose evolutionary history spans a time period that encompasses both deep divergences and more recent events during the Neogene (Smith *et al.*, 2017). Likewise, our improved homology search and filtering approaches aid in identifying the phylogenetic locations of polyploidy events. Moreover, we consider multiple lines of evidence for pinpointing the phylogenetic locations of polyploidy events, including orthogroup tree topology, Ks plots, and chromosome counts. These approaches identified 26 polyploidy events across Caryophyllales, include 10 newly reported and 16 previously identified (Yang *et al.*, 2015; Walker, *et al.*, 2017 **[Author, please insert either 'a' or 'b' to signify the correct Walker *et al.* (2017) citation]**). Importantly, two of these 26 events are suggested to be allopolyploidy events.

Species trees based on transcriptome data are concordant with previous analyses

The species trees we recovered are highly concordant with previous analyses of family-level relationships (Figs 7, S3, S4; Cuénoud *et al.*, 2002; Brockington *et al.*, 2009; Schäferhoff *et al.*, 2009; Arakaki *et al.*, 2011; Yang *et al.*, 2015). As seen in previous analyses, the placements of Sarcobataceae and Stegnospermataceae remain poorly supported despite using hundreds of loci. We found weak support for two nodes that had previously received high support using a small

number of loci. Previous analyses using plastome recovered Cactaceae as being sister to Portulacaceae with 100% bootstrap support (Arakaki *et al.*, 2011), but we recovered Anacampserotaceae+Portulacaceae as sisters to Cactaceae (modified phylome ICA = 0.31 and ASTRAL multi-locus bootstrap = 100). Previous studies recovered strong to moderate support for the monophyly of Portulacineae+Molluginaceae (likelihood bootstrap = 100, Arakaki *et al.*, 2011; Bayesian posterior probability = 0.94 and parsimony bootstrap < 50, Nyffeler & Eggli, 2010), but we found low support for the relationship (ICA = 0.06 and multi-locus bootstrap = 93). This confirms that while individual loci can be informative, there is a large amount of phylogenetic conflict among gene trees (Smith *et al.*, 2015; Walker *et al.*, 2017 **[Author, please insert either 'a' or 'b' to signify the correct Walker *et al.* (2017) citation]**). Future studies should dissect these cases of discordance using a gene-by-gene approach (Arcila *et al.*, 2017; Brown & Thomson, 2017; Shen *et al.*, 2017; Walker *et al.*, 2017 **[Author, please insert either 'a' or 'b' to signify the correct Walker *et al.* (2017) citation]**).

Many polyploidy events are associated with taxonomic units and/or habitat shifts

A notable pattern emerged showing that many polyploidy events occurred on branches leading to major taxa and/or involved clear habitat shifts (Fig. 7). For example, PHYT1 is located on the branch representing a transition from trees and large shrubs in wetter environments within the Neotropics to a radiation of arid- and semiarid-adapted herbs and subshrubs recognized as Tribe Nyctagineae of Nyctaginaceae (Douglas & Manos, 2007; Douglas & Spellenberg, 2010). Similarly, PORT1 at the base of Portulacineae is associated with the evolution of succulence (Nyffeler *et al.*, 2008; Edwards & Ogburn, 2012; Ogburn & Edwards, 2013). Additional polyploidy events are inferred along the branch leading to Polygonaceae (Schuster *et al.*, 2013) and the branch leading to Droseraceae, a carnivorous family (Rivadavia *et al.*, 2003).

Similar cases of polyploidy events at or near the base of major clade origins include seed plants (Jiao *et al.*, 2011), angiosperms (Jiao *et al.*, 2011), monocots (Jiao *et al.*, 2014), early eudicots (Jiao *et al.*, 2012), and Asteraceae (Barker *et al.*, 2016; Huang *et al.*, 2016). This hints at a potential relationship between genome duplication and evolutionary innovations (Soltis & Soltis, 2016). On the other hand, however, branches leading to major recognizable taxonomic units tend to be relatively long and thus had more time to accumulate changes. Hence, while correlations

between polyploidy and evolutionary novelty are intriguing, we must be cautious in assuming that polyploidy is the cause of such innovation (Smith *et al.*, 2017).

Inferring allopolyploidy from transcriptome data

Methods of polyploidy detection developed for genomes or low-copy nuclear loci are inadequate for datasets with isoforms and missing duplicated copies (Lott *et al.*, 2009; Jones *et al.*, 2013; Marcussen *et al.*, 2015; Thomas *et al.*, 2017). While we applied a stringent filter to minimize missing taxa in orthogroups (no more than two missing), differential gene loss or silencing following polyploidy events remained a problem. Given our goal of accurately pinpointing the phylogenetic location of polyploidy events and searching for allopolyploidy is highly dependent on taxon sampling, we only explored cases when Ks vs orthogroup tree-based mapping disagreed with each other.

Two allopolyploidy events were inferred in this study. We inferred the AMAR1 (Ks 0.4–0.65; Fig. 4) paleopolyploidy event followed by a nested, more recent polyploidy event (AMAR2) together were responsible for the observed Ks peaks, instead of a single, deeper event as previously reconstructed (Yang *et al.*, 2015). *Schiedea* also has a complex history (Fig. 5). While the polyploid origin of *Schiedea* was previously identified (Kapralov *et al.*, 2009; Yang *et al.*, 2015), by including its close relatives, *Honckenya* and *Scleranthus*, we show that all three species each had their own lineage-specific polyploidy event. *Schiedea* likely had a parental lineage other than the lineage leading to *Honckenya* (see schematic phylogram in Fig. 5). The putative allopolyploid origin of *Schiedea* adds to a growing list of Hawaiian endemic radiations with similar putative allopolyploid origins (Barrier *et al*., 1999; Yang & Berry, 2011; Marcussen *et al*., 2012; Roy *et al*., 2015), and demonstrates the importance of increased transcriptomic taxon sampling. Moving forward genome and transcriptome data will be essential for investigating selection, homeolog expression, gene silencing and loss in contributing to these divergence events following allopolyploidy.

Improved homology inference methods improve polyploidy mapping

In the original phylome approach (Huerta-Cepas *et al.*, 2011), each sequence from a seed species was used to search against a database of sequenced genomes. The resulting homologous sequences

were filtered, aligned, and phylogenetic trees were constructed. In this study, we made three modifications to the original phylome approach to enhance the ability to accommodate transcriptome data (Fig. S1). First, we merged putative homolog groups that represent gene duplication within Caryophyllales, ensuring that our final orthogroups are non-overlapping. Second, given the presence of multiple transcript isoforms and the inherent incompleteness of transcriptome datasets, we used transcriptome sequences as queries to search against the beet proteome for sorting transcriptome-derived sequences into backbone orthogroups constructed with genomes only. Lastly, to clean up spurious tips and isoforms, we added tip-trimming and long-branch cutting procedures. By taking this two-step, baited approach we were able to process a large number of taxa without going through the time consuming all-by-all homology search required by OrthoMCL (Li *et al.*, 2003) and OrthoFinder (Emms & Kelly, 2015). A second advantage of this modified phylome approach is that it avoids a graph-based clustering step, and hence is not biased by sequence length or phylogenetic relatedness among taxa (Emms & Kelly, 2015). The modified phylome approach is more effective than other baited approach such as HaMStR (Ebersberger *et al.*, 2009) in that it explicitly takes gene tree topology into account in distinguish ortholog from paralogs. However, because both the original phylome and the modified methodology start with a seed genome, the resulting orthogroup set is dependent on the quality and gene content of the focal genome.

In addition to the modified phylome approach, to overcome phylogenetic uncertainty associated with deep time scales we employed a second homology inference strategy that inferred subclade species trees using all-by-all homology search, Markov Clustering (van Dongen, 2000) of filtered hits, followed by alignment and tree trimming (Yang & Smith, 2014). We use the original Markov Clustering (MCL) instead of software packages like OrthoMCL (Li *et al.*, 2003) and OrthoFinder (Emms & Kelly, 2015) that aim at directly obtaining orthogroups using filtered and normalized BLAST hits. The normalization procedures used by these software packages were based on genome-derived data and were yet to be evaluated using transcriptome datasets that had isoforms, missing data, and assembly errors. By using the original MCL with relatively low inflation value (i.e. coarse clusters) and taking advantage of outgroup information to root and extract orthogroups, we were able to minimize the loss of gene duplication information in our dataset.

Our techniques found that for each inferred polyploidy event, approximately one-third of genes show clear evidence of duplication (i.e. they retain at least two overlapping taxa between paralogs), similar to the numbers identified in both transcriptomes and genomes (Yang *et al.*, 2015). For example, the PHYT2, AMAR1, and AMAR3 events follow this 'one-third rule' (Figs 2–6). When there is phylogenetic uncertainty, gene duplication events may be mapped to two adjacent nodes, each with lower percentages, such as observed for PORT1 and CARY1. Percentages of gene duplication can be inflated during rapid diversifications, where short internodes and phylogenetic uncertainty make it difficult to distinguish paralogous copies from isoforms using tree topology. Such inflated percentages of gene duplication can be seen at the base of Cactaceae and *Silene* (without polyploidy), and at PHYT1 and NCORE1 (following a polyploidy event).

Moving forward, additional taxon sampling of genomes and transcriptomes will be essential to identify additional polyploidy events and pinpoint their phylogenetic locations. Understanding the legacy of ancient polyploidy events in plant macroevolution will require many other forms of improved data as well, including functional studies of traits, molecular pathways, and genes themselves. Only then will we have a more comprehensive functional framework for understanding differential gene retention and diploidization following polyploidy events, and how polyploidy is linked to character evolution and niche shifts.

Acknowledgements

The authors thank H. Flores Olvera, H. Ochoterena, N. Douglas, S. Lavergne, T. Stoughton, L. Crawford, G. Kadereit, R. Puente, L. Majure, M. Howard, D. Anderson, M. Palmer, and K. Thiele for assisting with obtaining plant materials; the Cambridge University Botanic Gardens, Bureau of Land Management, White Sands Missile Range, US Forest Service, California State Parks, Rancho Santa Ana Botanic Garden, Desert Botanical Garden, Millennium Seed Bank, and Booderee National Park for granting access to their plant materials; M. R. M. Marchán-Rivadeneira and M. Parks for help with laboratory work; L. Y. Chen for help revising the manuscript; and constructive suggestions from three anonymous reviews. The molecular work of this study was conducted in part in the Genomic Diversity Laboratory at the University of

Michigan. Support came from the University of Michigan, Oberlin College, the National Geographic Society, the US National Science Foundation (DEB 1054539, DEB 1352907, and DEB 1354048), and a Natural Environment Research Council Independent Research Fellowship (NERC RG69516).

Author contributions

S.A.S., S.F.B., M.J.M., and Y.Y. designed research; M.J.M., Y.Y., J.M., J.O., S.F.B., and J.F.W. collected samples and carried out lab work; Y.Y. analyzed data and led the writing. All authors edited the manuscript and approved the final version.

References

Arakaki M, Christin PA, Nyffeler R, Lendel A, Eggli U, Ogburn RM, Spriggs E, Moore MJ, Edwards EJ. 2011. Contemporaneous and recent radiations of the world's major succulent plant lineages. *Proceedings of the National Academy of Sciences, USA* **108**(20): 8379–8384.

Arcila D, Ortí G, Vari R, Armbruster JW, Stiassny MLJ, Ko KD, Sabaj MH, Lundberg J, Revell LJ, Betancur-R R. 2017. Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. *Nature Ecology & Evolution* **1**: 0020.

Barker MS, Li Z, Kidder TI, Reardon CR, Lai Z, Oliveira LO, Scascitelli M, Rieseberg LH. 2016. Most Compositae (Asteraceae) are descendants of a paleohexaploid and all share a paleotetraploid ancestor with the Calyceraceae. *American Journal of Botany* **103**(7): 1203–1211.

Barrier M, Baldwin BG, Robichaux RH, Purugganan MD. 1999. Interspecific hybrid ancestry of a plant adaptive radiation: allopolyploidy of the Hawaiian silversword alliance (Asteraceae) inferred from floral homeotic gene duplications. *Molecular Biology and Evolution* **16**(8): 1105–1113.

Bell CD, Soltis DE, Soltis PS. 2010. The age and diversification of the angiosperms re-revisited. *American Journal of Botany* **97**(8): 1296–1303.

Brockington SF, Alexandre R, Ramdial J, Moore MJ, Crawley S, Dhingra A, Hilu K, Soltis DE, Soltis PS. 2009. Phylogeny of the Caryophyllales *sensu lato*: revisiting hypotheses

on pollination biology and perianth differentiation in the core Caryophyllales. *International Journal of Plant Sciences* **170**(5): 627–643.

Brockington SF, Yang Y, Gandia-Herrero F, Covshoff S, Hibberd JM, Sage RF, Wong GKS, Moore MJ, Smith SA. 2015. Lineage-specific gene radiations underlie the evolution of novel betalain pigmentation in Caryophyllales. *New Phytologist* **207**(4): 1170–1180.

Brown JM, Thomson RC. 2017. Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. *Systematic Biology* **66**(4): 517–530.

Byng JW, Chase MW, Christenhusz MJM, Fay MF, Judd WS, Mabberley DJ, Sennikov AN, Soltis DE, Soltis PS, Stevens PF *et al*. 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society* **181**(1): 1–20.

Cannon SB, McKain MR, Harkess A, Nelson MN, Dash S, Deyholos MK, Peng Y, Joyce B, Stewart CN, Rolf M. 2015. Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes. *Molecular Biology and Evolution* **32**(1): 193–210.

Caperta AD, Castro S, Loureiro J, Róis AS, Conceição S, Costa J, Rhazi L, Espírito Santo D, Arsénio P. 2016. Biogeographical, ecological and ploidy variation in related asexual and sexual *Limonium* taxa (Plumbaginaceae). *Botanical Journal of the Linnean Society* **183**(1): 75–93.

Chester M, Riley RK, Soltis PS, Soltis DE. 2015. Patterns of chromosomal variation in natural populations of the neoallotetraploid *Tragopogon mirus* (Asteraceae). *Heredity* **114**(3): 309–317.

Cuénoud P, Savolainen V, Chatrou LW, Powell M, Grayer ReJ, Chase MW. 2002. Molecular phylogenetics of Caryophyllales based on nuclear 18S rDNA and plastid *rbcL*, *atpB*, and *matK* DNA sequences. *American Journal of Botany* **89**(1): 132–144.

Dohm JC, Lange C, Holtgräwe D, Sörensen TR, Borchardt D, Schulz B, Lehrach H, Weisshaar B, Himmelbauer H. 2012. Palaeohexaploid ancestry for Caryophyllales inferred from extensive gene-based physical and genetic mapping of the sugar beet genome (*Beta vulgaris*). *Plant Journal* **70**(3): 528–540.

Dohm JC, Minoche AE, Holtgrawe D, Capella-Gutierrez S, Zakrzewski F, Tafer H, Rupp O, Sorensen TR, Stracke R, Reinhardt R *et al*. 2014. The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). *Nature* **505**(7484): 546–549.

Douglas N, Spellenberg R. 2010. A new tribal classification of Nyctaginaceae. *Taxon* **59**(3): 905–910.

Douglas NA, Manos PS. 2007. Molecular phylogeny of Nyctaginaceae: taxonomy, biogeography, and characters associated with a radiation of xerophytic genera in North America. *American Journal of Botany* **94**(5): 856–872.

Ebersberger I, Strauss S, von Haeseler A. 2009. HaMStR: profile hidden markov model based search for orthologs in ESTs. *BMC Evolutionary Biology* **9**(1): 157.

Edger PP, Heidel-Fischer HM, Bekaert M, Rota J, Glöckner G, Platts AE, Heckel DG, Der JP, Wafula EK, Tang M *et al*. 2015. The butterfly plant arms-race escalated by gene and genome duplications. *Proceedings of the National Academy of Sciences, USA* **112**(27): 8362–8366.

Edwards EJ, Ogburn RM. 2012. Angiosperm responses to a low-CO_2 World: CAM and C_4 photosynthesis as parallel evolutionary trajectories. *International Journal of Plant Sciences* **173**(6): 724–733.

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

Estep MC, McKain MR, Vela Diaz D, Zhong J, Hodge JG, Hodkinson TR, Layton DJ, Malcomber ST, Pasquet R, Kellogg EA. 2014. Allopolyploidy, diversification, and the Miocene grassland expansion. *Proc Natl Acad Sci U S A* **111**(42): 15149–15154.

Fawcett J, Maere S, Van de Peer Y. 2009. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proc Natl Acad Sci, USA* **106**: 5737–5742.

Fishman L, Willis JH, Wu CA, Lee YW. 2014. Comparative linkage maps suggest that fission, not polyploidy, underlies near-doubling of chromosome number within monkeyflowers (*Mimulus*; Phrymaceae). *Heredity* **112**(5): 562–568.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next generation sequencing data. *Bioinformatics* **28**(23): 3150–3152.

Glick L, Mayrose I. 2014. ChromEvol: assessing the pattern of chromosome number evolution and the inference of polyploidy along a phylogeny. *Mol Biol Evol* **31**(7): 1914–1922.

Hodgins KA, Lai Z, Oliveira LO, Still DW, Scascitelli M, Barker MS, Kane NC, Dempewolf H, Kozik A, Kesseli RV *et al*. 2014. Genomics of Compositae crops: reference transcriptome assemblies and evidence of hybridization with wild relatives. *Molecular Ecology Resources* **14**(1): 166–177.

Huang CH, Zhang C, Liu M, Hu Y, Gao T, Qi J, Ma H. 2016. Multiple polyploidization events across Asteraceae with two nested events in the early history revealed by nuclear phylogenomics. *Mol Biol Evol* **33**(11): 2820–2835.

Huerta-Cepas J, Capella-Gutierrez S, Pryszcz LP, Denisov I, Kormes D, Marcet-Houben M, Gabaldon T. 2011. PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic Acids Research* **39**: D556–D560.

Jaillon O, Aury J, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C *et al*. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**: 463–467.

Jiao Y, Leebens-Mack J, Ayyampalayam S, Bowers J, McKain M, McNeal J, Rolf M, Ruzicka D, Wafula E, Wickett N *et al*. 2012. A genome triplication associated with early diversification of the core eudicots. *Genome Biology* **13**(1): R3.

Jiao Y, Li J, Tang H, Paterson AH. 2014. Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. *Plant Cell* **26**(7): 2792–2802.

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang HY, Soltis PS *et al*. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**(7345): 97–100.

Jones G, Sagitov S, Oxelman B. 2013. Statistical inference of allopolyploid species networks in the presence of incomplete lineage sorting. *Syst Biol* **62**(3): 467–478.

Kane NC, King MG, Barker MS, Raduski A, Karrenberg S, Yatabe Y, Knapp SJ, Rieseberg LH. 2009. Comparative genomic and population genetic analyses indicate highly porous

genomes and high levels of gene flow between divergent *Helianthus* species. *Evolution* **63**(8): 2061–2075.

Kapralov MV, Stift M, Filatov DA. 2009. Evolution of genome size in Hawaiian endemic genus *Schiedea* (Caryophyllaceae). *Tropical Plant Biology* **2**(2): 77–83.

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

Kellogg EA. 2016. Has the connection between polyploidy and diversification actually been tested? *Current Opinion In Plant Biology* **30**: 25–32.

Kobert K, Salichos L, Rokas A, Stamatakis A. 2016. Computing the internode certainty and related measures from partial gene trees. *Mol Biol Evol* **33**(6): 1606–1617.

Lai Z, Kane NC, Kozik A, Hodgins KA, Dlugosch KM, Barker MS, Matvienko M, Yu Q, Turner KG, Pearl SA *et al*. 2012. Genomics of Compositae weeds: EST libraries, microarrays, and evidence of introgression. *American Journal of Botany* **99**(2): 209–218.

Li L, Stoeckert CJ J, Roos D. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**: 2178–2189.

Li Z, Baniaga AE, Sessa EB, Scascitelli M, Graham SW, Rieseberg LH, Barker MS. 2015. Early genome duplications in conifers and other seed plants. *Science Advances* **1**(10): e1501084.

Lohaus R, Van de Peer Y. 2016. Of dups and dinos: evolution at the K/Pg boundary. *Curr Opin Plant Biol* **30**: 62–69.

Lott M, Spillner A, Huber KT, Moulton V. 2009. PADRE: a package for analyzing and displaying reticulate evolution. *Bioinformatics* **25**(9): 1199–1200.

Löytynoja A, Goldman N. 2010. webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. *BMC Bioinformatics* **11**(1): 579.

Mandakova T, Li Z, Barker MS, Lysak MA. 2017. Diverse genome organization following 13 independent mesopolyploid events in Brassicaceae contrasts with convergent patterns of gene retention. *Plant J* **91**(1): 3–21.

Marcet-Houben M, Gabaldon T. 2015. Beyond the whole-genome duplication: phylogenetic evidence for an ancient interspecies hybridization in the Baker's yeast lineage. *PLoS Biol* **13**(8): e1002220.

Marcussen T, Heier L, Brysting AK, Oxelman B, Jakobsen KS. 2015. From gene trees to a dated allopolyploid network: insights from the angiosperm genus *Viola* (Violaceae). *Systematic Biology* **64**(1): 84–101.

Marcussen T, Jakobsen KS, Danihelka Jô, Ballard HE, Blaxland K, Brysting AK, Oxelman B. 2012. Inferring species networks from gene trees in high-polyploid North American and Hawaiian violets (*Viola*, Violaceae). *Systematic Biology* **61**(1): 107–126.

Mayrose I, Barker MS, Otto SP. 2010. Probabilistic models of chromosome number evolution and the inference of polyploidy. *Systematic Biology* **59**(2): 132–144.

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

Mirarab S, Nguyen N, Warnow T 2014. PASTA: ultra-large multiple sequence alignment. In: Sharan R, ed. *RECOMB 2014, LNBI 8394.* Basel, Switzerland: Springer International Publishing, 177–191.

Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* **30**(17): i541–i548.

Mirarab S, Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* **31**(12): 44–52.

Moore MJ, Soltis PS, Bell CD, Burleigh JG, Soltis DE. 2010. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proceedings of the National Academy of Sciences, USA* **107**(10): 4623–4628.

Nyffeler R, Eggli U. 2010. Disintegrating Portulacaceae: a new familial classification of the suborder Portulacineae (Caryophyllales) based on molecular and morphological data. *Taxon* **59**(1): 227–240.

Nyffeler R, Eggli U, Ogburn M, Edwards E. 2008. Variations on a theme: repeated evolution of succulent life forms in the Portulacineae (Caryophyllales). *Haseltonia*(14): 26–36.

Ogburn RM, Edwards EJ. 2013. Repeated origin of three-dimensional leaf venation releases constraints on the evolution of succulence in plants. *Current Biology* **23**(8): 722–726.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2, approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**(3): e9490.

Rivadavia F, Kondo K, Kato M, Hasebe M. 2003. Phylogeny of the sundews, *Drosera* (Droseraceae), based on chloroplast *rbcL* and nuclear 18S ribosomal DNA sequences. *American Journal of Botany* **90**(1): 123–130.

Rognes T. 2011. Faster Smith–Waterman database searches with inter-sequence SIMD parallelisation. *BMC Bioinformatics* **12**(1): 221.

Roy T, Cole LW, Chang T-H, Lindqvist C. 2015. Untangling reticulate evolutionary relationships among New World and Hawaiian mints (Stachydeae, Lamiaceae). *Molecular Phylogenetics and Evolution* **89**: 46–62.

Salichos L, Stamatakis A, Rokas A. 2014. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. *Molecular Biology and Evolution* **31**(5): 1261–1271.

Schäferhoff B, Müller KF, Borsch T. 2009. Caryophyllales phylogenetics: disentangling Phytolaccaceae and Molluginaceae and description of Microteaceae as a new isolated family. *Willdenowia* **39**(2): 209–228.

Schuster TM, Setaro SD, Kron KA. 2013. Age estimates for the buckwheat family Polygonaceae based on sequence data calibrated by fossils and with a focus on the amphi-Pacific *Muehlenbeckia*. *PLoS ONE* **8**(4): e61261.

Seo T-K. 2008. Calculating bootstrap probabilities of phylogeny using multilocus sequence data. *Mol Biol Evol.* **25**(5): 960–971.

Seo T-K, Kishino H, Thorne JL. 2005. Incorporating gene-specific variation when inferring and evaluating optimal evolutionary tree topologies from multilocus sequence data. *Proceedings of the National Academy of Sciences, USA* **102**(12): 4436–4441.

Shen X-X, Hittinger CT, Rokas A. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecology & Evolution* **1**: 0126.

Smith SA, Dunn CW. 2008. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. *Bioinformatics* **24**(5): 715–716.

Smith SA, Moore MJ, Brown JW, Yang Y. 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evolutionary Biology* **15:150**.

Smith SA, Brown JW, Yang Y, Brockington SF, Bruenn R, Drummond CP, Walker JF, Last N, Douglas, NA, Moore MJ. 2017. Disparity, diversity, and duplications in the Caryophyllales. *New Phytologist.*

Soltis PS, Marchant DB, Van de Peer Y, Soltis DE. 2015. Polyploidy and genome evolution in plants. *Current Opinion in Genetics & Development* **35**: 119–125.

Soltis PS, Soltis DE. 2016. Ancient WGD events as drivers of key innovations in angiosperms. *Current Opinion In Plant Biology* **30**: 159–165.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**(9): 1312–1313.

Stamatakis A 2016. *The RAxML v8.2.X manual.* [WWW document] URL https://github.com/stamatak/standard-RAxML/tree/master/manual [accessed 4 March 2017].

Steige KA, Slotte T. 2016. Genomic legacies of the progenitors and the evolutionary consequences of allopolyploidy. *Curr Opin Plant Biol* **30**: 88–93.

Tank DC, Eastman JM, Pennell MW, Soltis PS, Soltis DE, Hinchliff CE, Brown JW, Sessa EB, Harmon LJ. 2015. Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. *New Phytologist* **207**(2): 454–467.

Thomas GWC, Ather SH, Hahn MW. 2017. Gene-tree reconciliation with MUL-trees to resolve polyploidy events. *Syst Biol*. doi: 10.1093/sysbio/syx1044. **[Author, please replace DOI with volume and page range if available.]**

Thulin M, Moore AJ, El-Seedi H, Larsson A, Christin P-A, Edwards EJ. 2016. Phylogeny and generic delimitation in Molluginaceae, new pigment data in Caryophyllales, and the new family Corbichoniaceae. *Taxon* **65**(4): 775–793.

van Dongen S. 2000. *Graph clustering by flow simulation.* PhD thesis, University of Utrecht, Utrecht, the Netherlands.

Vanneste K, Baele G, Maere S, Van de Peer Y. 2014a. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous–Paleogene boundary. *Genome Res* **24**(8): 1334–1347.

Vanneste K, Maere S, Van de Peer Y. 2014b. Tangled up in two: a burst of genome duplications at the end of the Cretaceous and the consequences for plant evolution. *Philos Trans R Soc Lond B Biol Sci* **369**: 20130353.

Vanneste K, Van de Peer Y, Maere S. 2013. Inference of genome duplications from age distributions revisited. *Molecular Biology and Evolution* **30**(1): 177–190.

Walker JF, Brown JW, Smith SA. 2017a. Site and gene-wise likelihoods unmask influential outliers in phylogenomic analyses. *bioRxiv* doi: https://doi.org/10.1101/115774.

Walker JF, Yang Y, Moore MJ, Mikenas J, Timoneda A, Brockington SF, Smith SA. 2017b. Widespread paleopolyploidy, gene tree conflict, and recalcitrant relationships among the carnivorous Caryophyllales. *American Journal of Botany* **104**(6): 858–867.

Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH. 2009. The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences, USA* **106**(33): 13875–13879.

Xiang Y, Huang CH, Hu Y, Wen J, Li S, Yi T, Chen H, Xiang J, Ma H. 2016. Evolution of Rosaceae fruit types based on nuclear phylogeny in the context of geological times and genome duplication. *Mol Biol Evol* **34**(2): 262–281.

Yagi M, Kosugi S, Hirakawa H, Ohmiya A, Tanase K, Harada T, Kishimoto K, Nakayama M, Ichimura K, Onozaki T *et al*. 2014. Sequence analysis of the genome of carnation (*Dianthus caryophyllus* L.). *DNA Research* **21**(3): 231–241.

Yang Y, Berry PE. 2011. Phylogenetics of the Chamaesyce clade (*Euphorbia*, Euphorbiaceae): reticulate evolution and long-distance dispersal in a prominent C_4 lineage. *American Journal of Botany* **98**: 1486–1503.

Yang Y, Moore MJ, Brockington SF, Soltis DE, Wong GK-S, Carpenter EJ, Zhang Y, Chen L, Yan Z, Xie Y *et al*. 2015. Dissecting molecular evolution in the highly diverse plant clade Caryophyllales using transcriptome sequencing. *Molecular Biology and Evolution* **32**(8): 2001–2014.

Yang Y, Moore MJ, Brockington SF, Timoneda A, Feng T, Marx H, Walker JF, Smith SA. 2017. An efficient field and laboratory workflow for plant phylotranscriptomic projects. *Applications in Plant Sciences* **5**(3): 1600128.

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.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information tab for this article:

Fig. S1 Workflow for the modified phylome approach.

Fig. S2 Maximum quartet support species tree (MQSST) from ASTRAL analysis of individual subclade orthologous gene trees.

Fig. S3 Phylogram from RAxML analysis of the concatenated 624-gene supermatrix from modified phylomes, with ICA scores on branches.

Fig. S4 Maximum quartet support species tree (MQSST) from ASTRAL analysis of 624 orthologous gene trees from modified phylomes.

Fig. S5 Proportion of orthogroups showing duplications filtered by local tree topology.

Fig. S6 Distribution of within-taxon synonymous distances (Ks) among paralogs gene pairs.

Table S1 Sources of data and settings for assembly and translation

Table S2 Information for the 43 newly sequenced transcriptomes

Methods S1 All-by-all homology search for subclade datasets.

Please note: Wiley Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.

Fig. 1 Scenarios of polyploidy events (Kellogg, 2016). Letters 'A–E' represent taxon names, followed by chromosome numbers with the base number 'x', and schematic Ks plots (*x*-axis are Ks values, and *y*-axis are number of paralogous gene pairs).

Fig. 2 Best tree from RAxML analysis of concatenated supermatrix of the phytolaccoid clade and Aizoaceae (PHYT). Percentage values above branches indicate proportion of orthogroups showing duplication filtered by bootstrap percentage. Internode certainty all (ICA) values are given below branches.

Fig. 3 Best tree from RAxML analysis of concatenated supermatrix of Portulacineae and Molluginaceae (PORT). Percentage values above branches indicate proportion of orthogroups showing duplication filtered by bootstrap percentage. Internode certainty all (ICA) values are given below branches. Selected Ks plots based on BLASTN hits are shown below trees. Ks values <0.01 are not shown.

Fig. 4 Best tree from RAxML analysis of concatenated supermatrix of Amaranthaceae and Chenopodiaceae (AMAR). Percentage values above branches indicate proportion of orthogroups showing duplication filtered by bootstrap percentage. Internode certainty all (ICA) values are given below branches. Selected Ks plots based on BLASTN hits are shown below trees. Ks values <0.01 are not shown.

Fig. 5 Best tree from RAxML analysis of concatenated supermatrix of Caryophyllaceae (CARY). Percentage values above branches indicate proportion of orthogroups showing duplication filtered by bootstrap percentage. Internode certainty all (ICA) values are given below branches. Selected Ks plots based on BLASTN hits are shown below trees. Ks values <0.01 are not shown.

Fig. 6 Best tree from RAxML analysis of concatenated supermatrix of the clade sister to rest of the Caryophyllales (NCORE). Percentage values above branches indicate proportion of orthogroups showing duplication filtered by bootstrap percentage. Internode certainty all (ICA) values are given below branches.

Fig. 7 Species tree inferred by RAxML analysis of the supermatrix from modified phylomes. Polyploidy events are labeled according to Figs 2–6. When orthogroup tree topology vs Ks plots place a polyploidy event to different branches due to either phylogenetic uncertainty or allopolyploidy, we placed the event at the most recent common ancestor of taxa that share a within-taxon Ks peak.

Table 1 Statistics for homology and orthology inference. PHYT, PORT, AMAR, CARY, and NCORE are subclades within Caryophyllales (see Fig. 7)

	Caryophyllales	PHYT	PORT	AMAR	CARY	NCORE
Data type	Peptides	Coding sequences (CDS)				
Homology inference	Modified phylome	All-by-all				
Orthology inference (Yang *et al.*, 2014)	Rooted ingroup	One-to-one orthologs				
Number of taxa (ingroup + outgroup)	175+40	45+4	29+3	37+2	31+3	31+6
Minimal number of taxa per ortholog	160	49	31	38	34	36
Supermatrix dimension taxa × loci (columns)	178 × 624 (215669)	49 × 152 (217033)	32 × 171 (230873)	39 × 315 (453842)	34 × 736 (1130082)	37 × 213 (325966)
Supermatrix gene/character	92.6%/80.1%	100%/92.5%	97.8%/86.5%	98.1%/87.2%	100%/92.2%	97.5%/87.8%

occupancy						
Minimal ingroup taxa for mapping gene duplication	n/a	43	27	35	29	29
No. of orthogroups used for mapping gene duplications	n/a	2843	3577	4713	6686	1649

n/a, not applicable **[Author, please confirm inserted text 'n/a, not applicable' is correct]**.

Figure 1

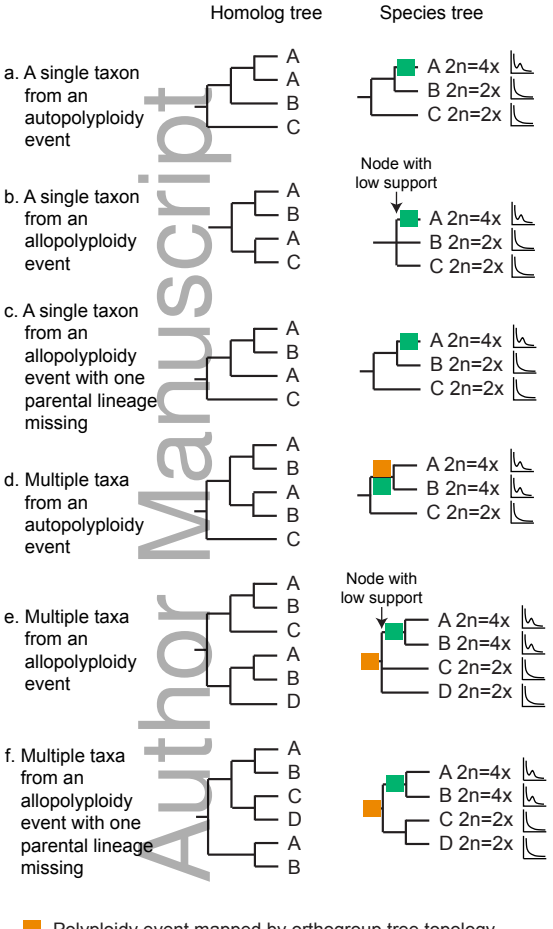

Figure 2

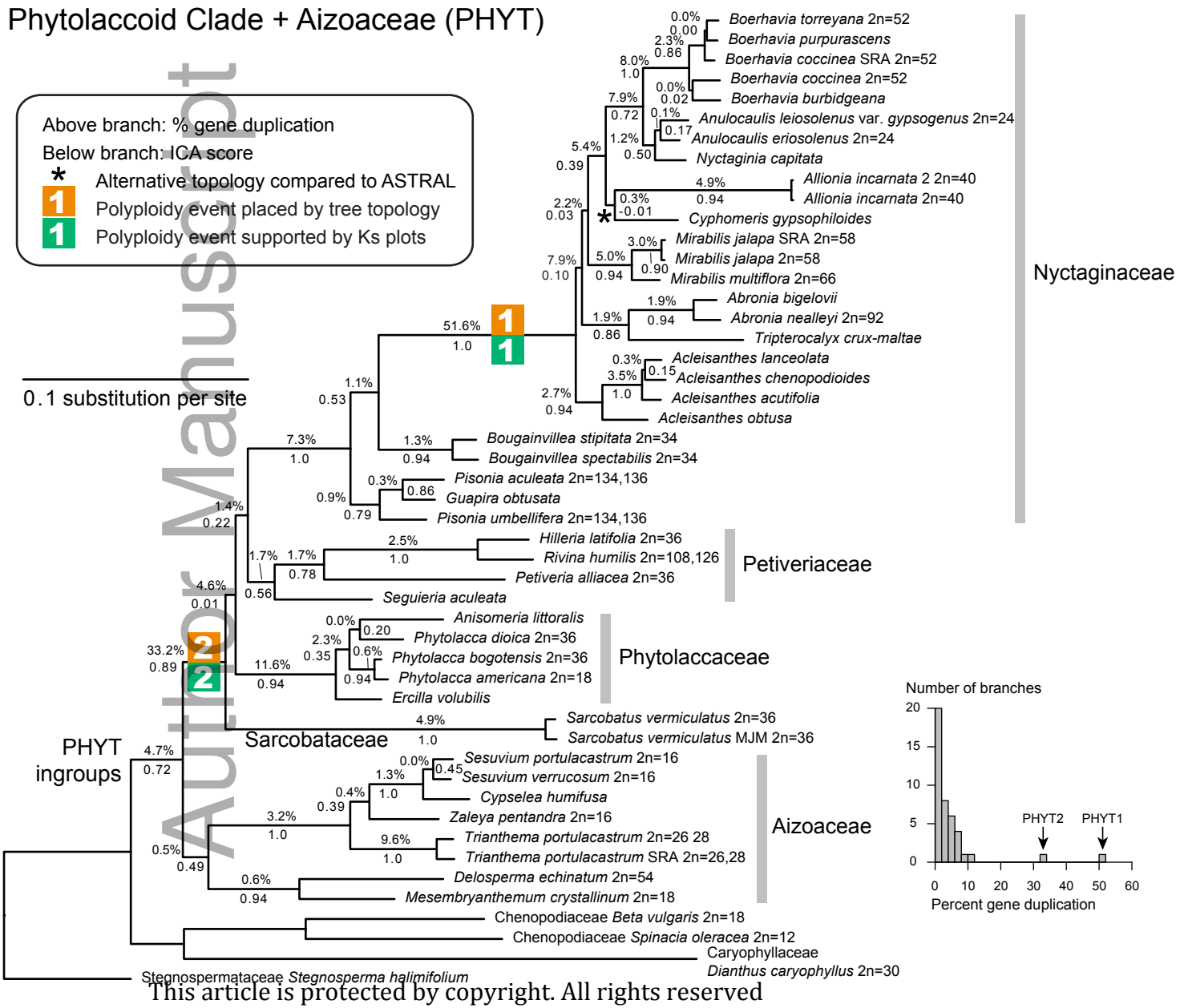

Figure 3

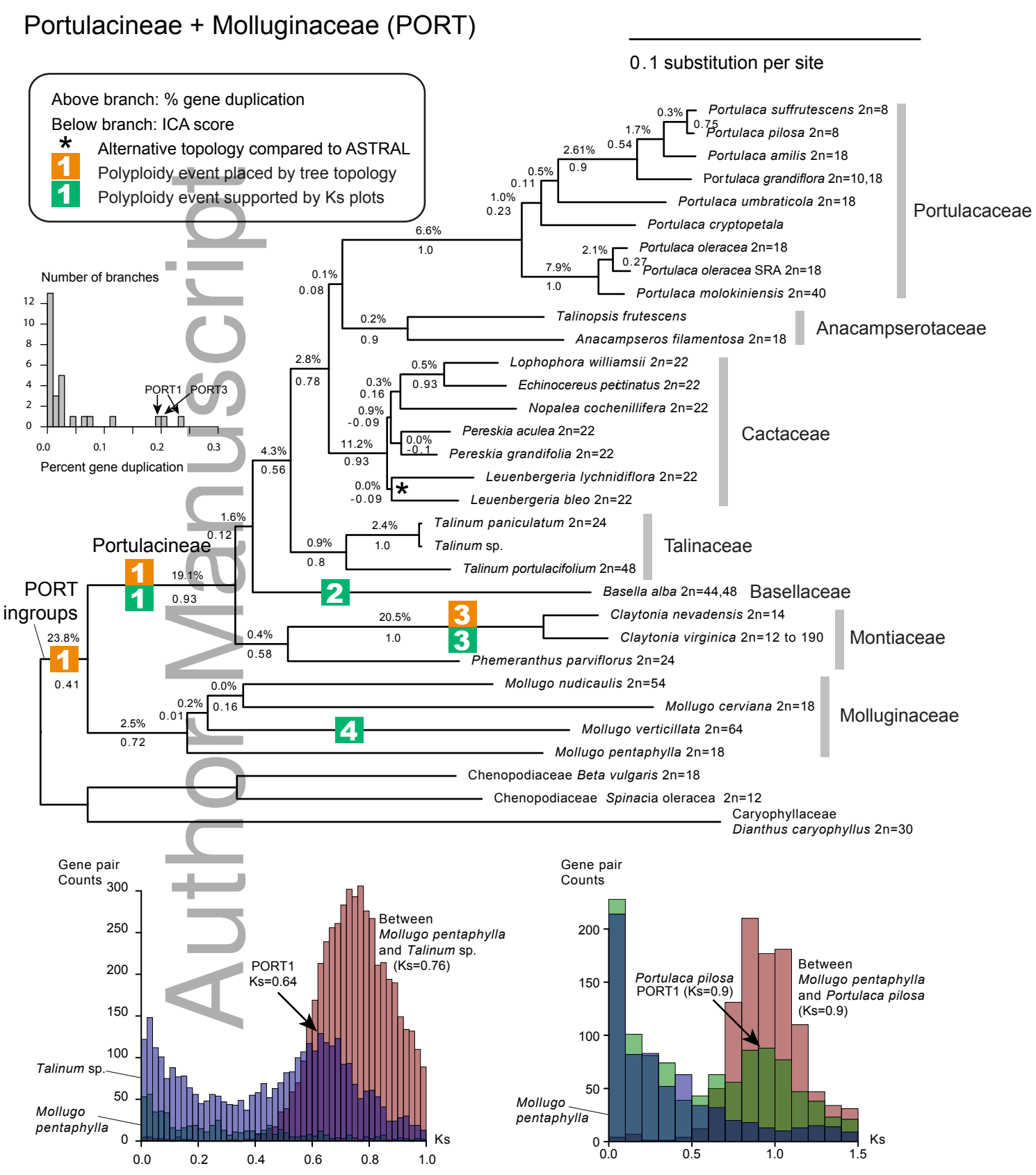

Figure 4

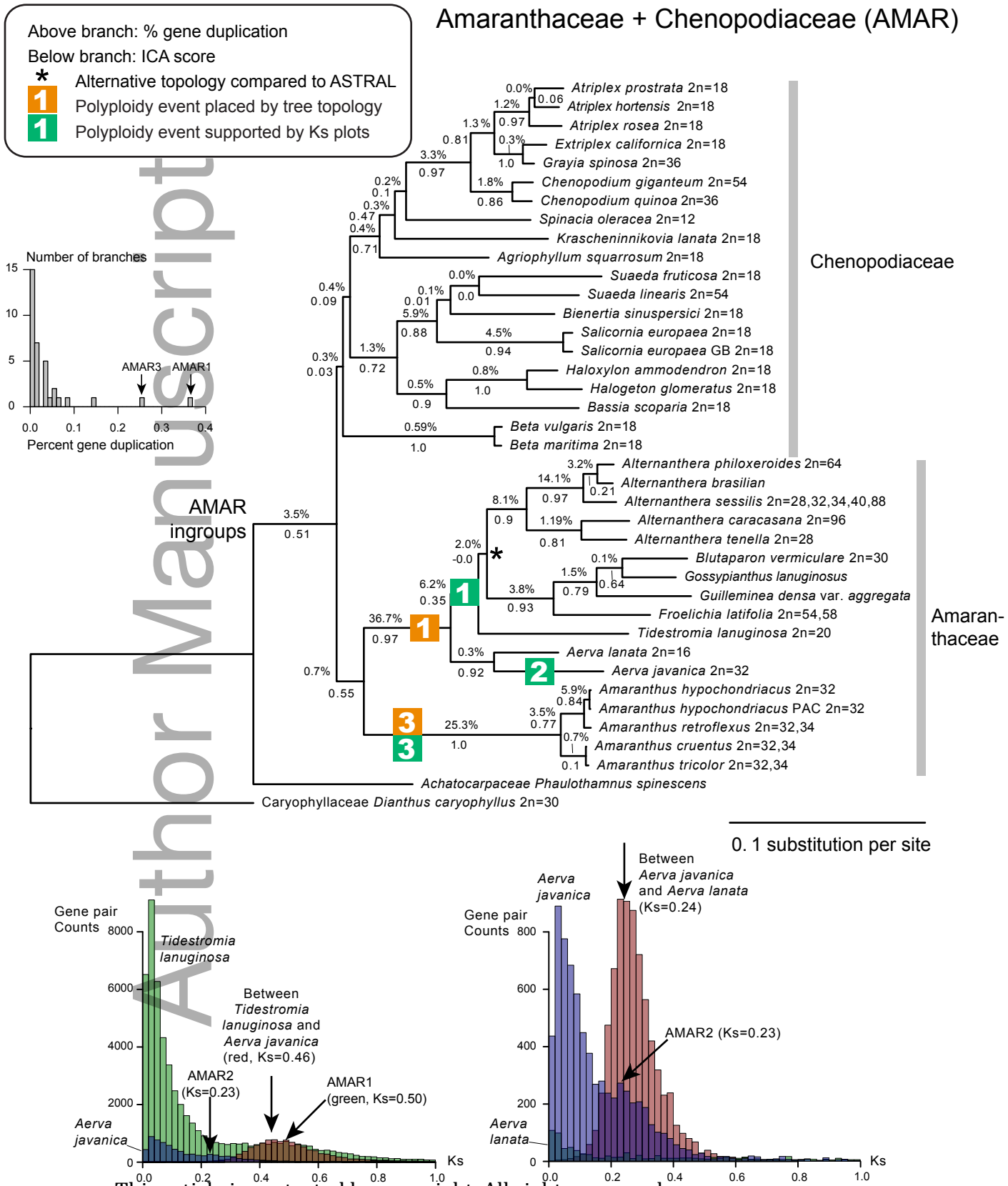

Figure 5

Caryophyllaceae (CARY)

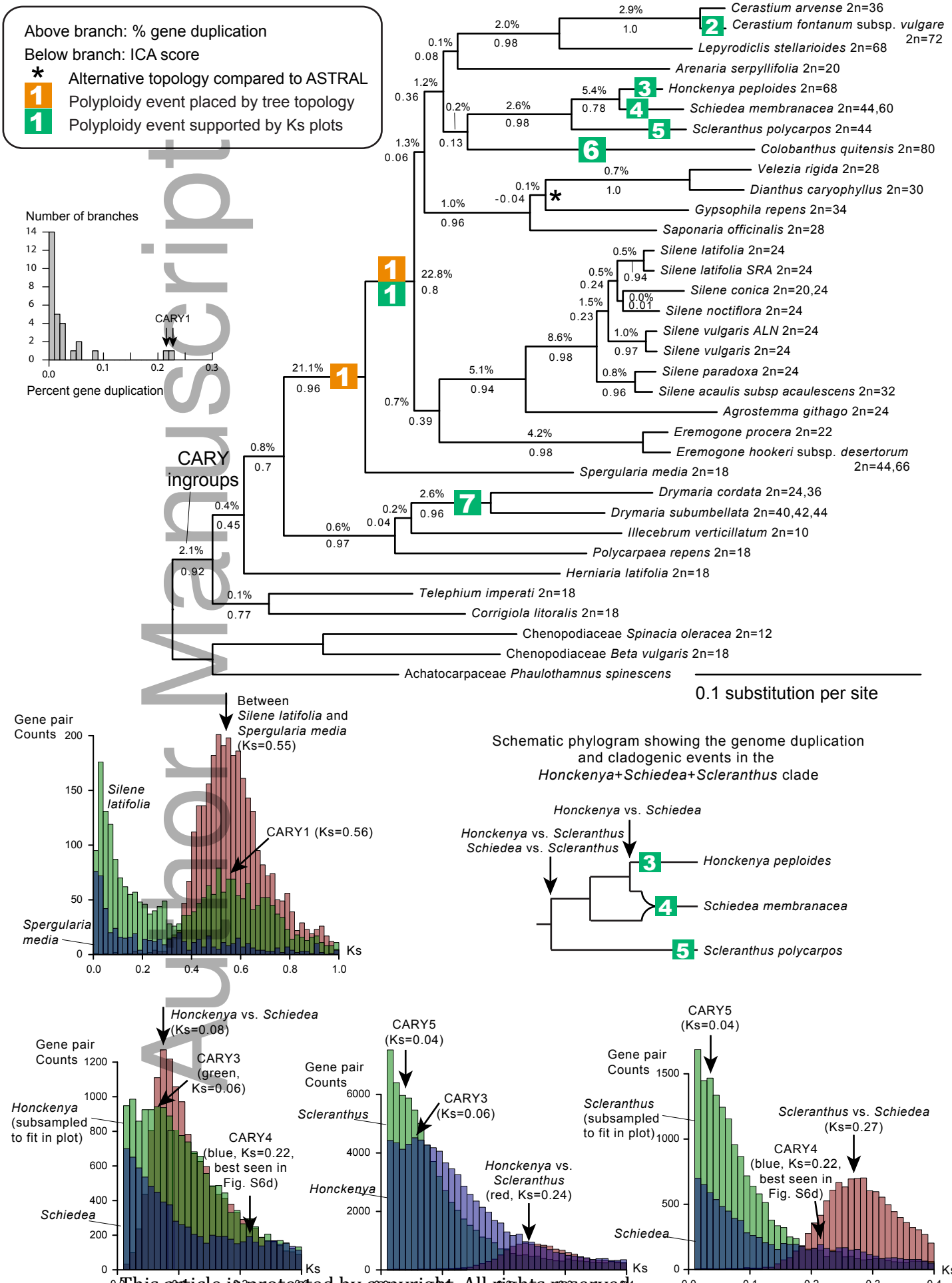

Figure 6

Clade sister to the rest of Caryophyllales (NCORE)

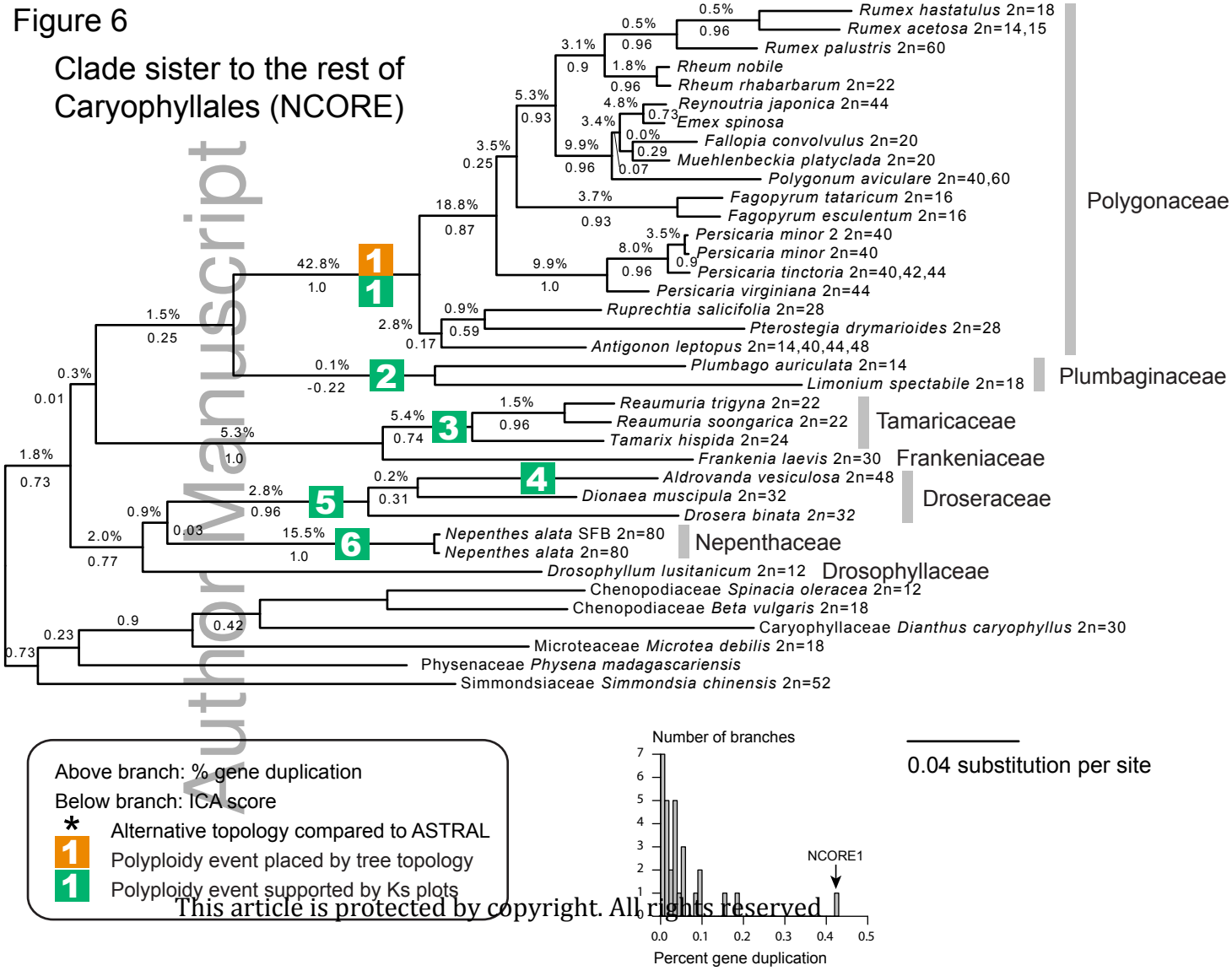

Figure 7