MBE Advance Access published April 2, 2015

Article, Discoveries

Dissecting molecular evolution in the highly diverse plant clade Caryophyllales using transcriptome sequencing

Ya Yang[*,1], Michael J. Moore[2], Samuel F. Brockington[3], Douglas E. Soltis[4,5,6], Gane Ka-Shu Wong[7,8,9], Eric J. Carpenter[7], Yong Zhang[9], Li Chen[9], Zhixiang Yan[9], Yinlong Xie[9], Rowan F. Sage[10], Sarah Covshoff[11], Julian M. Hibberd[11], Matthew N. Nelson[12], 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, Science Center K111, 119 Woodland St., Oberlin, Ohio 44074-1097 USA

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

[4]Department of Biology, University of Florida, Gainesville, FL 32611-8525, USA

[5]Florida Museum of Natural History, University of Florida, Gainesville, FL 32611-7800, USA

[6]Genetics Institute, University of Florida, Gainesville, FL 32610, USA

[7]Department of Biological Sciences, University of Alberta, Edmonton AB, T6G 2E9, Canada

[8]Department of Medicine, University of Alberta, Edmonton AB, T6G 2E1, Canada

[9]BGI-Shenzhen, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China

[10]Department of Ecology and Evolutionary Biology, University of Toronto, 25 Willcocks Street, Toronto, Ontario M5S 3B2, Canada

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

[12]School of Plant Biology, The University of Western Australia, 35 Stirling Highway, Crawley, Western Australia 6009, Australia

[*] Corresponding authors. Email: yangya@umich.edu or eebsmith@umich.edu

Abstract

Many phylogenomic studies based on transcriptomes have been limited to "single-copy" genes due to methodological challenges in homology and orthology inferences. Only a relatively small number of studies have explored analyses beyond reconstructing species relationships. We sampled 69 transcriptomes in the hyperdiverse plant clade Caryophyllales and 27 outgroups from annotated genomes across eudicots. Using a combined similarity- and phylogenetic tree-based approach, we recovered 10,960 homolog groups, where each was represented by at least eight ingroup taxa. By decomposing these homolog trees, and taking gene duplications into account, we obtained 17,273 ortholog groups, where each was represented by at least ten ingroup taxa. We reconstructed the species phylogeny using a 1,122-gene data set with a gene occupancy of 92.1%. From the homolog trees we found that both synonymous and nonsynonymous substitution rates in herbaceous lineages are up to three times as fast as in their woody relatives. This is the first time such a pattern has been shown across thousands of nuclear genes with dense taxon sampling. We also pinpointed regions of the Caryophyllales tree that were characterized by relatively high frequencies of gene duplication, including three previously unrecognized whole genome duplications. By further combining information from homolog tree topology and synonymous distance between paralog pairs, phylogenetic locations for 13 putative genome duplication events were identified. Genes that experienced the greatest gene family expansion were concentrated among those involved in signal transduction and oxidoreduction, including a cytochrome P450 gene that encodes a key enzyme in the betalain synthesis pathway. Our approach demonstrates a new approach for functional phylogenomic analysis in non-model species that is based on homolog groups in addition to inferred ortholog groups.

Keywords: Caryophyllales; substitution rate heterogeneity; paleopolyploidy; RNA-seq

Introduction

Transcriptome sequencing, or RNA-seq, has shown huge potential for understanding the genetic and genomic bases of diversification in non-model systems (for example, Barker *et al.* 2008; Dunn *et al.* 2008; Lee *et al.* 2011; Wickett *et al.* 2011; Delaux *et al.* 2014; Li *et al.* 2014; Misof *et al.* 2014; Sveinsson *et al.* 2014; Wickett *et al.* 2014; Cannon *et al.* 2015; Hollister *et al.* 2015). Previous phylogenomic studies based on transcriptomes have been limited by availability of data sets that include both high numbers of genes and dense taxon sampling. Methodological issues in homology and orthology inference, especially in accommodating the frequent genome duplications in plants, have resulted in the discarding of a large proportion of genes from previous phylogenomic studies (Yang and Smith 2014). These limitations, together with the dynamic nature of gene expression, gene duplication and loss, lineage specific heterogeneity in substitution rates and gene tree topology discordance have resulted in sparse matrices among nuclear genes in prior analyses, leading researchers to reduce data sets to a small number of genes for analysis. These challenges have limited many RNA-seq phylogenomic studies to inferring a species tree and only a limited number of studies have explored transcriptome-wide functional analyses beyond one-to-one orthologs or genes involved in a particular functional category (Barker *et al.* 2008; Lee *et al.* 2011).

We use recently developed tree-based homology inference methods (Yang and Smith 2014) that overcome many of the previous analytical limitations to illustrate the rich content of transcriptome data sets. We demonstrate our approach in the plant clade Caryophyllales by applying these methods to a data set of 69 transcriptomes and 27 genomes. With an estimated 11,510 species in 34 families (APG III; Bremer *et al.* 2009), the Caryophyllales represent approximately 6% of angiosperm species diversity, and are estimated to have a crown age of ca. 67–121 Ma (Bell *et al.* 2010; Moore *et al.* 2010). Species of the Caryophyllales are found across all continents and in all terrestrial ecosystems and exhibit extreme life history diversity, ranging from tropical trees to temperate annual herbs, and from long-lived desert succulents such as cacti to a diverse array of carnivorous plants [e.g., sundews (*Drosera*) and Old World pitcher plants (*Nepenthes*)]. The group also contains several independent origins of C_4 and CAM photosynthesis and exhibits repeated adaptation to warm, dry and high salinity environments (Sage *et al.* 2011; Edwards and Ogburn 2012; Kadereit *et al.* 2012). The extraordinary diversity in growth forms and ecological adaptations makes Caryophyllales an ideal group for investigating gene and genome evolution and heterogeneity in molecular substitution rate.

We apply a tree-based homology and orthology inference approach (Yang and Smith 2014) to examine phylogenetics, molecular substitution rate, and gene and genome duplications in the Caryophyllales. We use the inferred ortholog groups to reconstruct the phylogenetic relationships among major clades of Caryophyllales and to evaluate the heterogeneity in phylogenetic signal among ortholog

groups. We then use homolog trees to explore patterns of substitution rate heterogeneity among lineages within the Caryophyllales. Previous studies have linked variations in among-lineage substitution rates to a wide range of factors, such as temperature, UV radiation, water limitation, woody vs. herbaceous habit, parasitism, speciation rate, generation time, metabolic rate and rates of mitosis (Barraclough and Savolainen 2001; Davies *et al.* 2004; Kay *et al.* 2006; Wright *et al.* 2006; Smith and Donoghue 2008; Goldie *et al.* 2010; Gaut *et al.* 2011; Buschiazzo *et al.* 2012; Bromham *et al.* 2013; Lanfear *et al.* 2013). However, these studies have been based on either a small number of genes and (typically) the use of introns, or on the estimation of absolute rates between a few distantly related species pairs (Yue *et al.* 2010; Buschiazzo *et al.* 2012). Studies utilizing nuclear gene sequences have also suffered from sparse taxon sampling, where large phylogenetic distances separate the ingroup from a small number of distant outgroups. In this study, we leverage our high gene content and more thorough taxon sampling to test whether substitution rate between woody and herbaceous lineages are uniform across thousands of genes throughout the Caryophyllales. We also map gene duplication events inferred from homolog trees onto the species tree, revealing a number of previously unknown putative genome duplication events. Finally, we discuss the functional categories of genes that experienced high levels of gene family expansion within the Caryophyllales.

Results and Discussion

Tree-based homology and orthology inferences

Our data set comprises 69 Caryophyllales transcriptomes and 27 eudicot genomes. We took advantage of the availability of fully annotated genomes in eudicots that provide high quality outgroups for rooting the Caryophyllales (Goodstein *et al.* 2012). The homology inference first used similarity scores (E-values from BLASTP) to infer putative homolog groups (Yang and Smith 2014). For each putative homolog group, a multiple sequence alignment and a phylogenetic tree were inferred. Spurious branches were then pruned, and a refined alignment and tree were re-estimated. From the refined tree, sequence isoforms that form monophyletic or paraphyletic tips were removed. The remaining trees are the homolog trees (Yang and Smith 2014). A homolog group represents a single gene or gene family, or a monophyletic clade within a larger gene family. A total of 10,960 homolog groups were obtained, each consisting of at least eight of the 69 Caryophyllales taxa.

To estimate a species tree, we further decomposed unrooted homolog trees from amino acid sequences by extracting clades rooted by the earliest diverging species in our data set, *Aquilegia coerulea* (see Methods), and separating paralogs by inferring the location of gene duplication and pruning the subclade with the smaller number of taxa (Fig. 1A; Yang and Smith 2014). In this manner we recovered 17,273 ortholog groups with ten or more Caryophyllales taxa. The taxon occupancy curve (number of

taxa per ortholog group ranked from high to low) was nearly straight (Fig. 2), indicating that a moderate number of ortholog groups have high taxon coverage. Because *Aquilegia coerulea* was used for rooting, it was not present in the ortholog groups. An ortholog group with full taxon coverage therefore consisted of 95 taxa.

Phylogenetic relationships in Caryophyllales

We employed two alternative strategies for estimating the species tree and evaluating conflicting phylogenetic signals among genes (i.e. ortholog groups). We first constructed concatenated supermatrices, estimated the maximum likelihood (ML) phylogeny using RAxML, partitioning by each gene (CA-ML; Stamatakis 2006), and evaluated conflicting phylogenetic signal among genes by jackknife subsampling 10% and 30% of total number of genes (CA-JK10 and CA-JK30; Yang and Smith 2014). We chose the values of 10% and 30% to test the sensitivity of the topology to different ratios of subsampling. In addition to jackknife subsampling, we also carried out fast bootstrap on the supermatrices for comparison (CA-BS). Alternatively, we searched for the Maximum Quartet Support Species Tree (MQSST) using ASTRAL (Mirarab *et al.* 2014), starting from ML trees estimated from individual ortholog groups. Tree uncertainty was then evaluated using a two-stage bootstrap procedure that first resampled genes and then resampled characters in each resampled gene (MQSST-BS; Seo *et al.* 2005; Seo 2008). By applying bootstrapping and jackknifing, two non-parametric strategies, we were able to avoid making any assumption on sources of conflict, and we were therefore able to accommodate conflicting phylogenetic signal from potential sources such as hybridization, incomplete lineage sorting, errors, redundancy and missing data.

We conducted both CA-ML and MQSST analyses on a larger, 1,122-gene data set that had at least 85 taxa and 150 columns after trimming, and a smaller, 209-gene data set that included ortholog alignments with at least 90 taxa and 150 columns after trimming. The concatenated 1,122-gene supermatrix had 504,850 amino acid characters, with gene occupancy of 92.1% and character occupancy of 78.7%. The 209-gene supermatrix had an aligned length of 87,082 characters, gene occupancy of 95.6% and character occupancy of 83.6%. Two taxa (*Plumbago auriculata* and *Pereskia aculeata*) had the lowest gene occupancies, occurring in 12.5% and 27.0% of genes in our 1,122-gene supermatrix, respectively. All remaining taxa were present in over 50% of genes, with the vast majority (83 taxa) present in over 90% of genes (Table S1). In the 209-gene supermatrix, 88 out of the total 95 taxa were present in over 90% of genes (Table S1). Our gene and character occupancies are comparable to typical phylogenetic matrices derived from targeted PCR (Cuénoud *et al.* 2002; Brockington *et al.* 2009; Soltis *et al.* 2011). Hence, our data set sampled the genome broadly while minimizing uncertainties introduced by missing data.

Topologies of both the CA-ML tree and the MQSST based on the 1,122- and 209-gene data sets were well supported overall and were highly congruent with one another and with previous phylogenetic analyses (Cuénoud *et al.* 2002; Brockington *et al.* 2009, 2011; Schäferhoff *et al.* 2009; Soltis *et al.* 2011; Ruhfel *et al.* 2014). Most nodes from the 1,122-gene data set and a majority of nodes from the 209-gene data set received 100% support from all analyses (Fig. S1). Only five Caryophyllales branches received CA-JK30 < 99%, and all five also received MQSST-BS < 90% (indicated by arrows in Fig. 3 and Fig. S1). Below we review nodes that were relatively poorly supported and/or were recovered in incongruent positions.

Sarcobatus vermiculatus (Sarcobataceae) of the phytolaccoid clade [which we define here as comprising Nyctaginaceae, Phytolaccaceae *s.l.* (i.e., including *Agdestis*), Rivinaceae, and Sarcobataceae; Stevens 2015] was recovered in conflicting positions among the four analyses (Fig. 3 and Fig. S1). From the 1,122-gene CA-ML analysis, *Sarcobatus* was sister to *Phytolacca* (Phytolaccaceae; 89% BS, 81% JK30, and 59% JK10), whereas the 1,122-gene MQSST (100% BS) and both the 209-gene CA-ML (80%, 61% and 50% respectively) and the 209-gene MQSST (89% BS) placed it sister to remaining phytolaccoids. Previous phylogenetic analyses of Caryophyllales have also placed *Sarcobatus* within a polyphyletic Phytolaccaceae with low support (Cuénoud *et al.* 2002; Brockington *et al.* 2009, 2011; Schäferhoff *et al.* 2009). Examination of individual homolog trees revealed that *Sarcobatus vermiculatus* possesses a fast substitution rate compared to its close relatives. Its placement varied among homolog groups and was often poorly supported, with short internodes and low bootstrap percentages. It is possible that additional taxon sampling from Phytolaccaceae would improve support and topological stability for the position of *Sarcobatus*, which represents the only genus (of two closely related species) of Sarcobataceae. In particular, the inclusion of the monotypic genus *Agdestis* (Phytolaccaceae *s.l.*) would be desirable, as this genus was recovered as sister to *Sarcobatus* with moderate support in previous analyses (Cuénoud *et al.* 2002; Schäferhoff *et al.* 2009; Brockington *et al.* 2011).

Four additional branches within Caryophyllales also had relatively low support (Fig. 3 and Fig. S1): (1) the branch uniting Nepenthaceae, Frankeniaceae, Polygonaceae, and Plumbaginaceae, which was topologically congruent with previous analyses (Cuénoud *et al.* 2002; Brockington *et al.* 2009, 2011; Soltis *et al.* 2011) but had moderate to low support values (100% CA-BS / 87% CA-JK30 / 80% CA-JK10 / 21% MQSST-BS from the 1,122-gene analysis and 95%/77%/54%/32% respectively from the 209-gene analysis), (2) the branch uniting Portulacineae with Molluginaceae (100%/95%/77%/87% and 100%/87%/73%/83%, respectively), which was otherwise highly to moderately supported in all previous analyses (Arakaki *et al.* 2011; Soltis *et al.* 2011; Brockington *et al.* 2009, 2011), (3) the branch uniting *Bassia* with the clade of *Chenopodium* + *Atriplex* (Amaranthaceae), which was recovered in each analysis with moderate support (100%/100%/97%/78% and 100%/96%/75%/51%, respectively), while previous

analyses using chloroplast markers recovered a moderately-supported alternative topology of (((*Chenopodium*,*Atriplex*),*Beta*),*Bassia*), and (4) the branch uniting the Caryophyllaceae genera *Schiedea, Silene, Saponaria*, and *Dianthus*, which was weakly supported (88%/67%/56%/66% and 72%/48%/42%/44%, respectively). The former two taxa formed a grade in both the CA-ML and MQSST trees, but switched positions among resampling replicates. Both topologies—(*Schiedea,*(*Silene,*(*Saponaria*, *Dianthus*))) and (*Silene,*(*Schiedea,*(*Saponaria*, *Dianthus*)))—were present in roughly equal numbers of resampling replicates. Previous studies using chloroplast markers, however, recovered a third topology of *Dianthus,*(*Schiedea*, *Silene*) with moderate bootstrap support (Brockington *et al.* 2011). For both (3) and (4), additional sampling would be needed to distinguish among possible explanations such as ancient or recent hybridization, versus alternative (yet unlikely) explanations such as contamination and mislabeled samples.

Among the non-Caryophyllales eudicots, the placements of *Eucalyptus grandis* and *Cucumis sativus* also had low support values and were in conflict with analyses using chloroplast genomes (Xi *et al.* 2012; Ruhfel *et al.* 2014) or genes from both nuclear and organellar genomes (Soltis *et al.* 2011). In addition, the placement of *Linum usitatissimum* was weakly supported, consistent with previous analyses (Soltis *et al.* 2011; Xi *et al.* 2012). Given our focus on Caryophyllales, however, further discussion of these non-Caryophyllales incongruencies lies outside the scope of this paper.

A number of processes may account for the relatively low support and/or topological differences observed among the abovementioned nodes, including noise, hybridization, incomplete lineage sorting and rapid radiation. Concatenation has been criticized for inconsistency when gene trees vary disproportionally in branch lengths (Kolaczkowski and Thornton 2004) and for recovering topologies that are not supported by any gene tree (for example, Salichos and Rokas 2013). However, coalescence-based methods can also suffer from model violation from both non-random missing data and low phylogenetic signal in individual gene trees (Gatesy and Springer 2014; Springer and Gatesy 2014). Complications such as potential hybridization and bias in orthology inference also impact the nature and distribution of the conflicting signal in typical transcriptome data sets and as of yet are unexplored. Therefore, we present the CA-ML and MQSST topologies here because they are derived from the two alternative strategies for species tree estimation that are currently most appropriate for this data set. Also, given the complicated nature of conflict and the processes generating the conflict, species tree uncertainty is best evaluated by non-parametric measurements.

Given the overall high support values, the topological congruencies in the Caryophyllales among results from all four analyses except one single taxon, and the lack of branch length information in MQSST, we use the tree from the 1,122-gene CA-ML as the species tree for subsequent analyses and presentation of results.

Substitution rates in woody versus herbaceous lineages

To test the null hypothesis that substitution rate heterogeneity among lineages is not associated with woody versus herbaceous habits, we carried out five woody/herbaceous sister pair comparisons (Fig. 3). We extracted rooted Caryophyllales clades that had at least eight taxa from individual homologous gene trees, resulting in a total of 14,165 clades. For each of the five comparisons from each of the extracted Caryophyllales clade, we estimated the synonymous and nonsynonomous branch lengths, assuming that the nonsynonymous (dN) to synonymous (dS) substitution rate ratio (dN/dS) was constant among sites and within each woody or herbaceous lineage, but varied freely between the woody and the herbaceous sister lineages. We calculated the substitution rate contrast on each sister pair as the difference in branch lengths between the woody and herbaceous lineage divided by the shorter of the two. We designated the rate contrast to be positive when the substitution rate of the herbaceous lineage was faster than the woody lineage, and negative otherwise (Smith and Donoghue 2008). We used this non-parametric approach to avoid the uncertainties introduced by ancestral state reconstruction and molecular dating, and to avoid complications from gene duplications and conflicting tree topologies in certain areas. In addition, by focusing on non-overlapping clades, we were able to avoid multiple tests (Smith and Donoghue 2008).

A random subset of 3000 sister pairs was analyzed for each of the five comparisons (except 722 for contrast E due to the small data set size of *Plumbago auriculata*; Fig. 4 and Table 1). All had more positive than negative contrast values among both synonymous and nonsynonymous sites, and all ten sign tests were highly significant (Table 1A). To assess a potential node density bias, we tested cases where the woody and herbaceous sister lineage had the same number of tips and found that the pattern holds (Table 1B; Hugall and Lee 2007). Assuming that the habit of the taxa sister to each contrast was the best approximation of the ancestral habit type, the pattern also holds regardless of whether the direction of the transition was from woody to herbaceous (contrasts A and C, Fig. 3), from herbaceous to woody (contrasts D and E), or uncertain (contrast B). The highest rate differences were found in contrast A (herbaceous side with a rate 3.1× as high as the woody side for both dN and dS). Contrast C and E both had only one taxon on each side of the comparison and had less pronounced differences in rate compared to contrasts A, B and D. This is likely related to the fact that contrasts A, B and D had more taxa than contrasts C and E, which reduced effects of both rate stochasticity and noise introduced by transcriptome sequencing and assembly.

We found strong evidence that substitution rates in herbaceous lineages are up to three times as fast as in their woody relatives. The effect is observed in both synonymous and nonsynonymous sites and is most likely driven by differences in mutation rate between woody and herbaceous plants. Our study is the first to confirm such patterns across a large proportion of protein-coding genes in the plant genome

using dense taxon sampling. Although the pattern generally holds true across the Caryophyllales, there is one notable exception. The robust shrub *Sarcobatus vermiculatus* exhibits a higher substitution rate and longer branch than any other woody taxon sampled in the phytolaccoid clade, comparable to or exceeding certain herbaceous lineages (Fig. 3). The cause for this elevated rate is unclear.

Ancient genome duplications

Of all the homolog trees, 4,420 contained at least one clade with 60 or more Caryophyllales taxa and at least one non-Caryophyllales taxon. From these homolog trees, 167 extracted Caryophyllales clades, 3,651 asterid clades and 246 rosid clades had average bootstrap support values of at least 80% and were used for mapping gene duplication events. By mapping rooted clades with both high taxon occupancies and high bootstrap support values to the species tree (Fig. 1b), we detected six branches with highly elevated frequencies of gene duplications (circles in Figs. 3, S2). The highest concentration of gene duplications occurred at the base of Brassicaceae (Fig. 3) where 56% of the genes retained two or more duplicated copies in at least two taxa. The high percentage of duplicate genes reflects the alpha and beta paleotetraploidy events of Brassicaceae (Bowers *et al.* 2003; de Smet *et al.* 2013). The second location of elevated gene duplication was found along the branch uniting *Medicago*, *Phaseolus* and *Glycine* where 37% of genes retained duplicated copies, again corresponding to a known tetraploidy event (Fawcett *et al.* 2009). Since we only recorded gene duplications that shared at least two tips on both sides after trimming tips as a positive control for the Caryophyllales, we had less power for detecting genome duplications involving only two terminal taxa. Still, there was strong signal for genome duplication along the branch leading to *Solanum* in asterids, where 23% of genes retained at least two duplicated copies in both *S. lycopersicum* and *S. tuberosum*, which corresponds to a known paleohexaploidy event (The Tomato Genome Consortium 2012). These three genome duplication events represent all of the previously recognized genome duplications that involved at least two taxa in our data set (Lee *et al.* 2013).

Using the same approach we detected three previously unrecognized paleopolyploid events in Caryophyllales, corresponding to three branches with highly elevated percentages of genes showing gene duplications (circles in Figs. 3 and S2): at the base of Nyctaginaceae tribe Nyctagineae (represented by *Allionia, Anulocaulis, Boerhavia,* and *Mirabilis*; 31% of genes retaining duplicates), at the base of the phytolaccoid clade (36%), and at the base of the clade within Amaranthaceae containing *Aerva, Alternanthera,* and *Blutaparon* (24%). All three paleopolyploidy events had percentages of gene duplications comparable to the single paleotetraploidy event in Fabaceae. Distribution of synonymous substitutions (Ks plots) also supported the presence and location of all three genome duplication events (peaks 1, 2 and 5 respectively; numbered squares in Figs. 3 and S2). Due to the up to three-fold synonymous rate heterogeneity between woody and herbaceous clades, peak 2 shifted its location

between ca. 0.3 to 1 (Fig. S2). In addition, when peak 1 was present, peak 2 partly overlapped with both peak 1 and the peak between 1.8 and 2 that represents the genome triplication event that occurred in early eudicots (Jiao *et al.* 2012). Typically, detection of genome duplications has been based on distribution of synonymous distances between paralog pairs, or synteny analyses when genomes are available (Fawcett *et al.* 2009; Jiao *et al.* 2011; Jiao *et al.* 2012; Lee *et al.* 2013; Sveinsson *et al.* 2014; Cannon *et al.* 2015). However, both rate heterogeneity and partially overlapping peaks are known to make detecting genome duplication using the distribution of Ks plots challenging (Vanneste *et al.* 2013). By incorporating both homolog tree topology and Ks plots, we were able to overcome both challenges and recover previously unrecognized ancient genome duplications.

Six additional peaks (4, 7, 9, 10, 12 and 13) were found, each involving a single terminal taxon (Figs. 3, S2). These were not detectable using the homolog tree topology as we required at least two terminals to be duplicated to identify a duplicated gene. The lack of a peak at a similar position in the Ks plots of their sister taxa indicate that these seven genome duplication events occurred along each terminal branch. The genome duplication event along the branch leading to *Schiedea* (peak 7) after its split from *Silene latifolia* is further supported by a previous Ks analysis based on EST (Kapralov *et al.* 2008). Similarly, both peaks 6 and 11 involve two taxa each and are therefore less detectable from the homolog tree topology alone after masking monophyletic and paraphyletic tips. This is especially true for *Plumbago auriculata*, which is an unusually low-coverage data set (Table S1). The lack of homolog groups showing duplication at the node uniting Plumbaginaceae and Polygonaceae supports that peaks 11 and 12 are from independent genome duplications that happened after this split. In addition, the lack of a Ks peak around 0.8 in *Frankenia* indicates that peak 13 is independent from peaks 11 and 12.

Detecting genome duplications using homolog tree topologies also has less power where ancient and rapid divergences occurred and when individual gene trees had little phylogenetic information. This is likely the case for peaks 3 (Portulacineae) and 8 (within Caryophyllaceae), both near nodes of low support values (Figs. 3, S2). In both cases, gene duplications found from homolog trees were mapped to the node coincident with the most recent common ancestor of taxa sharing the Ks peak, as well as deeper nodes. Further analyses and additional taxon sampling in Caryophyllaceae are needed to distinguish among possible explanations such as lack of information, incomplete lineage sorting, and/or allopolyploidy.

Earlier phylogenetic analyses have suggested that the backbone of Caryophyllales may have radiated rapidly (e.g., Brockington *et al.*, 2009). Nevertheless, the absence of genome duplications along the deeper branches of Caryophyllales is supported by analyses of the *Beta vulgaris* genome (Dohm *et al.* 2012; 2014). Comparisons of this genome with the grape genome showed one-to-one synteny and no

evidence of lineage-specific genome duplication along the branches leading to *Beta vulgaris* (Dohm *et al.* 2012; 2014).

We detected 13 genome duplication events within the Caryophyllales. Our methods differ from previous studies that detect genome duplication using transcriptome data (Barker *et al.* 2008; Sveinsson *et al.* 2014; Cannon *et al.* 2015) in that we use homolog tree topology much more extensively to infer the precise phylogenetic locations of peaks in Ks distributions. Although one can apply sophisticated models to overcome saturation and partially overlapping peaks, and to correct for substitution rate heterogeneity among lineages (Vanneste *et al.* 2013), phylogenetic tree inference incorporates information beyond Ks distances and therefore is capable of overcoming these challenges more effectively.

Exploring functional categories of genes

Aside from the phylogenetic locations of gene duplications, we also investigated functional categories of genes that showed high taxon occupancy or lineage-specific duplications. We carried out a gene ontology (GO) enrichment analysis for the 4,420 homolog groups that included at least 60 Caryophyllales taxa against a background of all 10,960 homolog groups with at least eight Caryophyllales taxa. The enriched GO terms involved a variety of housekeeping functions such as aminoacyl-tRNA ligase activity, protein targeting to vacuole, Golgi vesicle transport, monosaccharide/oligosaccharide/polysaccharide metabolic process, carbohydrate biosynthetic process, proteolysis, cellular amino acid metabolic process, nucleotide biosynthetic process, response to metal ion, and chloroplast organization (Table S2). The fact that homolog groups with a high number of taxa are enriched for housekeeping functions confirms that these genes are constitutively expressed and are successfully recovered by our homology inference method.

Next, we investigated gene duplicates that can be identified based on phylogenetic trees with the current species sampling (Table 2). It is important to note two potential issues that likely lead to some level of underestimation of gene duplications in our data. First, our transcriptome data were restricted to certain tissue types at specific developmental stages (mostly leaves), and hence could not capture all gene expression throughout the plant. Second, the fact that *de novo* assembly often cannot reliably distinguish recent paralogs from allelic variation led to our exclusion of monophyletic and paraphyletic tips that belonged to the same taxon in the homolog trees. This likely also obscured some recent gene duplication events, especially in areas with less taxon sampling. Despite the possible underestimation of the true number of gene duplications, we postulate that if we can detect clear examples of genes that have undergone lineage-specific duplication within the Caryophyllales—i.e., if we can recover copies that are co-expressed at levels high enough to detect using *de novo* assembly—then these duplicated copies may contribute to lineage-specific adaptation.

We ranked extracted Caryophyllales clades with at least five taxa by either the highest number of tips for any single taxon (Table 2A) or the total number of tips (Table 2B) that can be identified based on our taxon sampling and homolog trees. Of the 13 Caryophyllales clades with the highest copy number for any taxon, five represented transposable elements (Table 2A). The remaining clades were involved in four functional classes: oxidoreductase (nucleoredoxin 1-like, cytochrome P450 and other oxidoreductase family proteins), pollen recognition (G-type lectin S-receptor-like serine/threonine-protein kinase), disease resistance signaling (NBS-LRR type resistance protein), and transcriptome factors (protein FAR1-related sequence 5-like). Most of the taxa with the highest number of tips in homolog trees belong to Nyctaginaceae, likely due to both the denser taxon sampling in Nyctaginaceae and the multiple genome duplication events in the phytolaccoid clade (Fig. 3).

To minimize the effect of taxon sampling, we ranked Caryophyllales clades by the total number of tips. Of the ten Caryophyllales clades with the highest total number of tips, three belonged to the cytochrome P450 family (Table 2B), with the largest clade consisting of 361 tips. Functional categories for these Caryophyllales clades include oxidoreductase, disease resistance, peptide transportation, kinases, and lipid degradation (Table 2B). Notably, the Caryophyllales clade that was annotated as "Cytochrome P450 71A25-like" encodes a key enzyme in the betalain synthesis pathway (Fig. S3; Hatlestad *et al.* 2012). Betalains are a group of pigments unique to core Caryophyllales that produce bright yellow, pink and violet colorations in vegetative, flower and fruit tissues, and that have been shown to have strong antioxidant activity (Gandía-Herrero and García-Carmona 2013). Further exploration of the locations of gene family expansion may be useful in identifying recruitment of key candidate genes for future functional dissection of other ecophysiological traits and developmental processes that are frequently encountered in Caryophyllales and are of general interest, such as photosynthetic pathways (Christin *et al.* 2014), salt tolerance, succulence and carnivory. Additional, carefully targeted taxon sampling will increase the power of such an approach.

Conclusion

This study demonstrates a new approach for functional phylogenomic analysis in non-model species using homolog groups in addition to inferred ortholog groups. We show that transcriptome data sets can provide high-quality homolog and ortholog sets that are rich resources not only for reconstructing phylogenies but also for mapping substitution rate shifts and gene duplications. The power and utility of these data sets are highly contingent on efficient and accurate clustering, aligning, and cleaning in order to obtain high-quality homolog trees. From these homolog trees we were able to recover candidate genes that experienced extensive lineage-specific gene family expansion, which provide a rich resource for investigating genes that may have contributed to adaptive changes.

Materials and Methods

Data availability

Raw reads for the six newly generated transcriptomes were deposited in the NCBI Sequence Read Archive (BioProject: PRJNA261527; see Table S3 for individual accessions). Assembled sequences, alignments, and trees are available from Dryad (doi:10.5061/dryad.33m48). Scripts used for analyses in this study are also available from Dryad with notes and updates available from https://bitbucket.org/yangya/caryophyllales_mbe_2015.

Sampling and laboratory procedures

Assembled transcripts and translated amino acid sequences from 66 Caryophyllales transcriptome data sets were obtained from the One Thousand Plants (1KP) Consortium (https://sites.google.com/a/ualberta.ca/onekp/). Each transcriptome was derived from leaves and/or flower buds. RNA isolation, quality control, library preparation and sequencing procedures were summarized in Johnson *et al.* (2012), and collection information is available on the 1KP website. After preliminary phylogenomic analysis and examination of gene trees and the matrix occupancy, four 1KP transcriptomes were removed due to contamination or small data set size (Table S1). Of the 62 1KP transcriptomes used in this study, *Bassia scoparia* was published by the 1KP pilot study labeled as its synonym *Kochia scoparia* (Matasci *et al*. 2014; Wickett *et al*. 2014). The remaining 61 transcriptomes from 1KP are published here for the first time (Table S1). In addition, the sugar beet (*Beta vulgaris*) EST data were downloaded from The *Beta vulgaris* Resource website (http://bvseq.molgen.mpg.de) (Herwig *et al.* 2002; Dohm *et al.* 2012).

We generated six additional transcriptomes for this study (Table S3). Young leaves and/or flower buds were frozen in liquid nitrogen and stored at -80°C. Total RNA was extracted using the BIO-RAD Aurum Total RNA Mini Kit (Life Science Research, Hercules, California) and quantified using the NanoDrop 1000 Spectrophotometer (Thermo Scientific, Wilmington, Delaware). Libraries were prepared using the TruSeq RNA Sample Preparation Kit v2 (Illumina, Inc., San Diego, California) with modifications (Supplementary Methods), and quantified on an Agilent 2100 Bioanalyzer (Agilent Technologies, Inc., Santa Clara, California). All six libraries were multiplexed and sequenced on one lane of an Illumina HiSeq2000 sequencer at the University of Michigan DNA Sequencing Core.

For outgroup comparison, non-redundant transcripts and amino acid sequences for 27 eudicot species derived from genome annotations were downloaded from Phytozome v9 (Goodstein *et al.* 2012).

Sequence processing

Paired end 101-bp reads from the six newly generated transcriptomes were trimmed and filtered using the following procedures: read pairs with either of the reads having average quality score ≤ 32 were removed; 3' end nucleotides with quality scores < 10 were trimmed; post-trimming read pairs with either of the reads ≤ 30 bases were removed; and read pairs with adaptor contamination were removed. The remaining reads were assembled using Trinity v2012-02-25 with default settings (Grabherr *et al.* 2011). Assembled reads were translated using TransDecoder v2012-08-15 with default settings (Haas *et al.* 2013). Translation of the beet EST sequences was carried out using FrameDP v1.2.0 (Gouzy *et al.* 2009) with default parameters, except for setting framed_maximum_peptide_length to 1500, and using a custom blast database consisting of amino acid sequences from the 27 eudicot genome annotations. All translated amino acid sequences from each of the 69 Caryophyllales transcriptome data sets were reduced with CD-HIT v4.6 (-c 0.99 -n 5) (Fu *et al.* 2012).

Homology inference

Initial homology searches were carried out using all-by-all BLASTP with an E-value cutoff of 10^{-5} and -max_target_seqs 100. BLASTP hits with identical matches > 50% and query coverage / query length > 0.7 were used for homology inference. Putative homolog groups were obtained using Markov clustering (MCL v12-068; van Dongen 2000) with an E-value cutoff of 10^{-30} and an inflation value of 1.4. Only sequences 40 amino acids or longer in length were retained, and only clusters with at least eight Caryophyllales taxa were retained.

Each cluster was aligned using MAFFT v7 (Katoh and Standley 2013) with "--genafpair --maxiterate 1000" ("--auto" for clusters with more than 1,000 sequences), and alignments were trimmed using Phyutility v2.2.6 (Smith and Dunn 2008) with "-clean 0.05". Phylogenetic trees were estimated using FastTree v2.1.7 (Price *et al.* 2009; 2010). Because trees with branches longer than 2 substitutions per site were most likely caused by either error or clusters being pulled together by conserved domains, such long branches were cut, the sequences on each subtree were realigned, the resulting matrix was trimmed, and the phylogeny was re-estimated following the same procedure as the initial round until no branch was longer than 2. This process was repeated with a more stringent branch length cutoff of 0.5, set by the distribution of branch lengths among the ingroups. Final alignment was estimated using SATé v2.2.7 with "iter-limit=3" (Liu *et al.* 2009; Liu *et al.* 2012). ML phylogenetic inference was carried out using RAxML v7.3.2 (Stamatakis 2006) with the model PROTCATWAG. Branches > 0.6 were cut; spurious terminal branches > 0.2 and more than 10 times longer than their sisters were also removed (Yang and Smith 2014). Finally, since the transcriptome assembly process produced multiple isoforms for each gene that formed monophyletic or paraphyletic groups, only the one with the highest number of

characters in trimmed alignments was retained as the representative, with the rest removed (Yang and Smith 2014). The remaining trees were the homolog trees.

Orthology inference and species tree estimation

Clades rooted by the basal eudicot *Aquilegia coerulea* (Ranunculaceae; this species was sister to all others in our analyses) were extracted from homolog trees. For each extracted clade, gene duplication events were inferred by the presence of one or more duplicated taxa between two sides; the side with the smaller number of taxa was subsequently pruned (Fig. 1A). This procedure was carried out from root to tips iteratively on all subclades until no taxon duplication was present. The alignment for each ortholog group was constructed by extracting aligned sequences from the homologs and trimming by Phyutility with "-clean 0.3" (Smith and Dunn, 2008). After visualizing taxon occupancy statistics of ortholog groups, a supermatrix was constructed by concatenating trimmed ortholog alignments with at least 90 taxa and 150 amino acids. A second supermatrix was constructed using ortholog groups with at least 85 taxa and at least 150 amino acids after trimming. A ML tree was estimated from each supermatrix using RAxML with the PROTCATWAG model, partitioning each ortholog group. Node support was evaluated by 200 fast bootstrap replicates. To explicitly evaluate the conflicting phylogenetic signal among ortholog groups, we also carried out 200 jackknife replicates for both supermatrices, randomly subsampling 30% and 10% of the total numbers of genes without replacement, while keeping each gene region intact (Yang and Smith 2014).

In addition to the concatenated analyses, we also searched for the Maximum Quartet Support Species Tree (MQSST) using ASTRAL v4.7.3 (Mirarab *et al.* 2014). A ML tree for each ortholog group that had at least 85 taxa and 150 characters after trimming was estimated using RAxML with the PROTCATWAG model. Phylogenetic uncertainty within each ortholog group was estimated using 200 fast bootstrap replicates in RAxML. Uncertainty for the MQSST was estimated with 100 bootstrap replicates using a two-stage multi-locus bootstrap strategy (Seo *et al.* 2005; Seo 2008) as implemented in ASTRAL. A second ASTRAL analysis was carried out with the same setting using ortholog groups that had at least 90 taxa and 150 characters after trimming.

Substitution rate contrasts

To investigate the relationship between habit (woody vs. herbaceous) and substitution rates, we first plotted the habit of all sampled Caryophyllales taxa to the species tree (Fig. 3). Although the cactus *Lophophora williamsii* is not a typical tree or shrub, it is similar to typical woody plants in having prominent and persistent above ground tissue and was therefore treated as woody.

We extracted rooted Caryophyllales clades with at least eight taxa from individual homologous gene trees. For each of these Caryophyllales clades, all corresponding amino acid sequences were pooled, and an alignment was estimated using PRANK v140110 (Löytynoja and Goldman 2010) with the default settings except that a single iteration was carried out using the extracted clade as the guide tree. Each resulting alignment was back translated using PAL2NAL v14 (Suyama *et al*. 2006), and the codon alignment was trimmed by requiring each codon column to have a minimum of 20% unambiguous codons. Both the extracted Caryophyllales clades and the trimmed codon alignments were used to estimate the synonymous (dS) and nonsynonomous (dN) substitution rates in CODEML, as part of the PAML package v4.8a (Yang 2007). For each of the five woody/herbaceous pairs in a given Caryophyllales clade, we labeled the woody clade as clade #1, its sister herbaceous clade as clade #2, and the rest of the tree as the "background". Ratios of dN/dS for the background, clade #1 and clade #2 were estimated separately with constant dN/dS across sites (model = 2, NSites = 0) and codon frequencies estimated by F3x4. The synonymous branch length contrasts at a given internal node were calculated by stepwise averaging of the corresponding synonymous branch lengths estimated by CODEML from the tips to the node, and dividing the difference in branch lengths between the two sister clades by the shorter of the two (Smith and Donoghue 2008). Similarly, the nonsynonymous contrast was calculated using the corresponding nonsynonomous branch lengths estimated by CODEML. For all woody/herbaceous sister pairs, if the herbaceous side had longer average branch lengths, we designated the contrast values to be positive; values were negative if the woody side had longer branch lengths.

Inferring the phylogenetic location of gene duplications

Homolog trees that contained clades with at least one outgroup and 60 ingroup taxa were subject to 100 rapid bootstrap replicates in RAxML to evaluate branch support. Caryophyllales clades containing at least 60 of the 69 taxa, asterid clades containing all three taxa and rosid clades containing at least 18 of the 22 taxa were extracted from homolog trees. Extracted clades with average bootstrap support values ≥ 80% were used for locating gene duplication events. A gene duplication event was recorded at a node on the extracted clade if the two branches from this node shared two or more taxa. Duplication events from each extracted clade were mapped to the corresponding node on the species tree. When the gene tree had missing taxa or had a conflicting topology compared to the species tree, we mapped the duplication event onto the corresponding most recent common ancestor on the species tree. When nested duplication events were detected from an extracted clade, only one duplication event was mapped to each corresponding node on the species tree for each extracted clade (Fig. 1B).

Distribution of synonymous distances (Ks) between paralogous gene pairs

To verify putative genome duplications detected from homolog tree topology, we visualized the distribution of Ks within each of the 69 Caryophyllales taxa. The recently published peptides and CDS from the *Beta vulgaris* genome annotation (Dohm *et al.* 2014) were used instead of the EST sequence for the Ks analysis. The 1KP accession HURS was used instead of KJAA (combined analysis of accession HURS and BZMI) for the Ks analysis due to the discovery that BZMI is a mixture of *Mollugo pentaphylla* and *M. verticillata*. For all 69 Caryophyllales transcriptome data sets, highly similar sequences were reduced with CD-HIT (-c 0.99 -n 5). An all-by-all BLASTP was carried out within each taxon using an E-value cutoff of 10 and -max_target_seq set to 20. Resulting hits with pident < 20% or niden < 50 amino acids were removed. Sequences with ten or more hits were removed to avoid over-representation of gene families that experienced multiple recent duplications. The remaining paralog pairs and their corresponding CDS were used to calculate Ks values using the pipeline https://github.com/tanghaibao/bio-pipeline/tree/master/synonymous_calculation (accessed November 29, 2014). The pipeline first carries out pairwise protein alignment using default parameters in ClustalW (Larkin *et al.* 2007), back-translates the alignment to a codon alignment using PAL2NAL, and calculates the synonymous substitution rate (Ks) using yn00 as part of the PAML package, with Nei-Gojobori correction for multiple substitutions (Nei and Gojobori 1986).

Functional annotation

We obtained the *Arabidopsis thaliana* gene that was most closely related to the Caryophyllales clade of interests based on homolog tree topology and used its locus ID to represent the Caryophyllales clade in the GO analysis. When multiple *A. thaliana* genes are present in the clade sister to the Caryophyllales, a random one was chosen. GO enrichment analysis was carried out using GOrilla (Eden *et al.* 2009) with control for False Discovery Rate (Benjamini and Hochberg 1995).

In addition to automated annotation for GO enrichment analysis, we also manually annotated clades having the most gene family expansion. Peptide alignments were visually examined, and a sequence with high completeness in the alignment was BLASTed against the non-redundant protein database in NCBI. Functional annotations of top BLAST hits were performed using the UniProt database (http://www.uniprot.org/).

Acknowledgements

We thank Jim Leebens-Mack, John Cheeseman, Michael Deyholos, Mark Chase, Neal Stewart and Dmitry Filatov for contributing RNA samples; Kew Gardens for access to the living collections; Marta Laskowski, Rich Cronn, Matt Parks, Tara Jennings, Liliana Cortés-Ortiz and Ingrid Jordon-Thaden for help with lab work; Joseph Brown and Cody Hinchliff for helpful discussions; and six anonymous

reviewers for help improving the manuscript. The molecular work of this study was conducted in the Genomic Diversity Laboratory of the Department of Ecology and Evolutionary Biology, University of Michigan. The 1000 Plants (1KP) initiative, led by G.K.S.W., is funded by the Alberta Ministry of Enterprise and Advanced Education, Alberta Innovates Technology Futures (AITF), Innovates Centre of Research Excellence (iCORE), Musea Ventures, and BGI-Shenzhen. Additional support came from a National Science Foundation award (DEB 1352907 and DEB 1354048) to S.A.S., M.J.M. and S.F.B., a MCubed initiative award to S.A.S., and a National Geographic Society award to M.J.M.

Author contributions

S.A.S., M.J.M., S.F.B. and Y.Y. designed research; G.K.S.W. and E.J.C. coordinated the 1KP initiative; S.C., M.J.M., S.F.B., Y.Y., R.F.S., J.M.H., D.E.S. and M.N.N. conducted lab work and contributed to the taxon sampling; Y.Z., L.C., Z.Y., Y.X., M.J.M. and Y.Y. performed library preparation, sequencing and sequence processing; Y.Y. and S.A.S. analyzed data; and Y.Y. and S.A.S. led the writing.

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. *Proc Natl Acad Sci U S A.* 108(20):8379–8384.

Barker MS, Kane NC, Matvienko M, Kozik A, Michelmore RW, Knapp SJ, Rieseberg LH. 2008. Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Mol Biol Evol.* (11):2445–2455.

Barraclough TG, Savolainen V. 2001. Evolutionary rates and species diversity in flowering plants. *Evolution* 55(4):677–683.

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

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57(1):289–300.

Bowers JE, Chapman BA, Rong J, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422(6930):433–438.

Bremer B, Bremer Kr, Chase M, Fay M, Reveal J, Soltis D, Soltis P, Stevens P. 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Botanical Journal of the Linnean Society* 161(2):105–121.

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, Walker RH, Glover BJ, Soltis PS, Soltis DE. 2011. Complex pigment evolution in the Caryophyllales. *New Phytologist* 190(4):854–864.

Bromham L, Cowman PF, Lanfear R. 2013. Parasitic plants have increased rates of molecular evolution across all three genomes. *BMC Evol. Biol.* 13(1):126.

Buschiazzo E, Ritland C, Bohlmann J, Ritland K. 2012. Slow but not low: genomic comparisons reveal slower evolutionary rate and higher dN/dS in conifers compared to angiosperms. *BMC Evol. Biol.* 12(1):8.

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. *Mol Biol Evol.* 32(1):193–210.

Christin P-A, Arakaki M, Osborne CP, Bräutigam A, Sage RF, Hibberd JM, Kelly S, Covshoff S, Wong GK-S, Hancock L et al. 2014. Shared origins of a key enzyme during the evolution of C_4 and CAM metabolism. *Journal of Experimental Botany* 65(13):3609–3621.

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.

Davies TJ, Savolainen V, Chase MW, Moat J, Barraclough TG. 2004. Environmental energy and evolutionary rates in flowering plants. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 271(1553):2195–2200.

Delaux P-M, Varala K, Edger PP, Coruzzi GM, Pires JC, Ané J-M. 2014. Comparative phylogenomics uncovers the impact of symbiotic associations on host genome evolution. *PLoS Genet* 10(7):e1004487.

de Smet R, Adams KL, Vandepoele K, Van Montagu MCE, Maere S, Van de Peer Y. 2013. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proc Natl Acad Sci U S A.* 110(8):2898–2903.

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*). *The 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.

Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD et al. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* 452(7188):745–749.

Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. 2009. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 10(1):48.

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.

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 U S A.* 106:5737–5742.

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.

Gandía-Herrero F, García-Carmona F. 2013. Biosynthesis of betalains: yellow and violet plant pigments. *Trends in Plant Science* 18(6):334–343.

Gatesy J, Springer MS. 2014. Phylogenetic analysis at deep timescales: Unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. *Molecular Phylogenetics and Evolution* 80(0):231–266.

Gaut B, Yang L, Takuno S, Eguiarte LE. 2011. The patterns and causes of variation in plant nucleotide substitution rates. *Annual Review of Ecology, Evolution, and Systematics* 42:245–266.

Goldie X, Gillman L, Crisp M, Wright S. 2010. Evolutionary speed limited by water in arid Australia. *Proceedings of the Royal Society B: Biological Sciences* 277(1694):2645–2653.

Goodstein DM, Shu SQ, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N et al. 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40(D1):D1178–D1186.

Gouzy J, Carrere S, Schiex T. 2009. FrameDP: sensitive peptide detection on noisy matured sequences. Bioinformatics 25(5):670–671.
Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* 29(7):644–654.
Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M et al. 2013. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* 8(8):1494–1512.
Hatlestad GJ, Sunnadeniya RM, Akhavan NA, Gonzalez A, Goldman IL, McGrath JM, Lloyd AM. 2012. The beet R locus encodes a new cytochrome P450 required for red betalain production. *Nat Genet* 44(7):816–820.
Herwig R, Schulz B, Weisshaar B, Hennig S, Steinfath M, Drungowski M, Stahl D, Wruck W, Menze A, O'Brien J. 2002. Construction of a 'unigene' cDNA clone set by oligonucleotide fingerprinting allows access to 25 000 potential sugar beet genes. *The Plant Journal* 32(5):845–857.
Hollister JD, Greiner S, Wang W, Wang J, Zhang Y, Wong GK-S, Wright SI, Johnson MTJ. 2015. Recurrent loss of sex is associated with accumulation of deleterious mutations in *Oenothera*. *Mol Biol Evol.* doi: 10.1093/molbev/msu345
Hugall AF, Lee MSY. 2007. The likelihood node density effect and consequences for evolutionary studies of molecular rates. *Evolution* 61(10):2293–2307.
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, 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.
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. doi:10.1186/gb-2012-1113-1181-r1183.
Johnson MTJ, Carpenter EJ, Tian Z, Bruskiewich R, Burris JN, Carrigan CT, Chase MW, Clarke ND, Covshoff S, dePamphilis CW et al. 2012. Evaluating methods for isolating total RNA and predicting the success of sequencing phylogenetically diverse plant transcriptomes. *PloS One* 7(11):e50226.
Kadereit G, Ackerly D, Pirie MD. 2012. A broader model for C_4 photosynthesis evolution in plants inferred from the goosefoot family (Chenopodiaceae s.s.). Proceedings of the Royal Society B: Biological Sciences 279(1741):3304–3311.
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. *Mol Biol Evol.* 30(4):772–780.
Kay K, Whittall J, Hodges S. 2006. A survey of nuclear ribosomal internal transcribed spacer substitution rates across angiosperms: an approximate molecular clock with life history effects. *BMC Evol. Biol.* 6(1):36.
Kolaczkowski B, Thornton JW. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* 431(7011):980–984.
Lanfear R, Ho SYW, Jonathan Davies T, Moles AT, Aarssen L, Swenson NG, Warman L, Zanne AE, Allen AP. 2013. Taller plants have lower rates of molecular evolution. *Nat Commun* 4:1879.
Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23(21):2947–2948.

Lee EK, Cibrian-Jaramillo A, Kolokotronis S-O, Katari MS, Stamatakis A, Ott M, Chiu JC, Little DP, Stevenson DW, McCombie WR et al. 2011. A functional phylogenomic view of the seed plants. *PLoS Genet* 7(12):e1002411.
Lee T-H, Tang H, Wang X, Paterson AH. 2013. PGDD: a database of gene and genome duplication in plants. *Nucleic Acids Res.* 41(D1):D1152–D1158.
Li F-W, Villarreal JC, Kelly S, Rothfels CJ, Melkonian M, Frangedakis E, Ruhsam M, Sigel EM, Der JP, Pittermann J et al. 2014. Horizontal transfer of an adaptive chimeric photoreceptor from bryophytes to ferns. *Proc Natl Acad Sci U S A.* 111(18):6672–6677.
Liu K, Raghavan S, Nelesen S, Linder CR, Warnow T. 2009. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science* 324(5934):1561–1564.
Liu K, Warnow TJ, Holder MT, Nelesen SM, Yu J, Stamatakis AP, Linder CR. 2012. SATé-II: Very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Systematic Biology* 61(1):90–106.
Löytynoja A, Goldman N. 2010. webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. *BMC Bioinformatics* 11(1):579.
Matasci N, Hung L-H, Yan Z, Carpenter E, Wickett N, Mirarab S, Nguyen N, Warnow T, Ayyampalayam S, Barker M et al. 2014. Data access for the 1,000 Plants (1KP) project. *GigaScience* 3(1):17.
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.
Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG. 2014. Phylogenomics resolves the timing and pattern of insect evolution. *Science* 346(6210):763–767.
Moore MJ, Soltis PS, Bell CD, Burleigh JG, Soltis DE. 2010. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proc Natl Acad Sci U S A.* 107(10):4623–4628.
Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol.* 3(5):418–426.
Price MN, Dehal PS, Arkin AP. 2009. FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol.* 26(7):1641–1650.
Price MN, Dehal PS, Arkin AP. 2010. FastTree 2, approximately maximum-likelihood trees for large alignments. *PloS One* 5(3):e9490.
Ruhfel B, Gitzendanner M, Soltis P, Soltis D, Burleigh J. 2014. From algae to angiosperms-inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC Evol. Biol.* 14(1):23.
Sage RF, Christin P-A, Edwards EJ. 2011. The C_4 plant lineages of planet earth. *Journal of Experimental Botany* 62(9):3155–3169.
Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497(7449):327–331.
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.
Seo T-K, Kishino H, Thorne JL. 2005. Incorporating gene-specific variation when inferring and evaluating optimal evolutionary tree topologies from multilocus sequence data. *Proc Natl Acad Sci U S A.* 102(12):4436–4441.
Seo T-K. 2008. Calculating bootstrap probabilities of phylogeny using multilocus sequence data. *Mol Biol Evol.* 25(5):960–971.
Smith SA, Donoghue MJ. 2008. Rates of molecular evolution are linked to life history in flowering plants. *Science* 322(5898):86–89.
Smith SA, Dunn CW. 2008. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. *Bioinformatics* 24(5):715–716.

Soltis DE, Smith SA, Cellinese N, Wurdack KJ, Tank DC, Brockington SF, Refulio-Rodriguez NF, Walker JB, Moore MJ, Carlsward BS et al. 2011. Angiosperm phylogeny: 17 genes, 640 taxa. *American Journal of Botany* 98(4):704–730.
Springer MS, Gatesy J. 2014. Land plant origins and coalescence confusion. *Trends in Plant Science* 19(5):267–269.
Stamatakis A. 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22(21):2688–2690.
Stevens PF. 2015. Angiosperm Phylogeny Website. Version 13. Available at http://www.mobot.org/mobot/research/apweb/.
Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 34(suppl 2):W609–W612.
Sveinsson S, McDill J, Wong GK, Li J, Li X, Deyholos MK, Cronk QC. 2014. Phylogenetic pinpointing of a paleopolyploidy event within the flax genus (*Linum*) using transcriptomics. *Annals of Botany* 113:753–761.
The Tomato Genome Consortium. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485(7400):635–641.
van Dongen S. 2000. Graph clustering by flow simulation. Ph.D thesis, University of Utrecht, Utrecht, Netherlands.
Vanneste K, Van de Peer Y, Maere S. 2013. Inference of genome duplications from age distributions revisited. *Mol Biol Evol.* 30(1):177–190.
Wickett NJ, Honaas LA, Wafula EK, Das M, Huang K, Wu B, Landherr L, Timko MP, Yoder J, Westwood JH et al. 2011. Transcriptomes of the parasitic plant family Orobanchaceae reveal surprising conservation of chlorophyll synthesis. *Current Biology* 21(24):2098–2104.
Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, Ayyampalayam S, Barker MS, Burleigh JG, Gitzendanner MA et al. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc Natl Acad Sci U S A.* 111(45):E4859–E4868.
Wright S, Keeling J, Gillman L. 2006. The road from Santa Rosalia: A faster tempo of evolution in tropical climates. *Proc Natl Acad Sci U S A.* 103(20):7718–7722.
Xi Z, Ruhfel BR, Schaefer H, Amorim AM, Sugumaran M, Wurdack KJ, Endress PK, Matthews ML, Stevens PF, Mathews S et al. 2012. Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. *Proc Natl Acad Sci U S A.* 109(43):17519–17524.
Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol.* 24(8):1586–1591.
Yang Y, Smith SA. 2014. Orthology inference in non-model organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. *Mol Biol Evol.* 31(11):3081–3092.
Yue J-X, Li J, Wang D, Araki H, Tian D, Yang S. 2010. Genome-wide investigation reveals high evolutionary rates in annual model plants. *BMC Plant Biology* 10(1):242.

Figure Legends

Figure 1. Schematic outline of analyses in this study. (A) Homology and orthology inferences and (B) mapping gene duplications detected from homolog trees to the species tree

Figure 2. Ortholog groups ranked from high to low by number of taxa represented. Only ortholog groups represented by at least ten Caryophyllales taxa were shown. An ortholog group with full taxon occupancy included 95 taxa.

Figure 3. Species tree from RAxML analysis of the 1,122-gene supermatrix. Numbers on branches indicate proportion of genes showing duplication; duplications were not investigated on unmarked branches.

Figure 4. Distribution of substitution rate contrasts. Contrast values were calculated from individual homologous gene trees for five woody-herbaceous sister pairs (A–E in Fig. 3). Rate contrasts were considered positive (blank) when the herbaceous (H) side possessed a faster rate than the woody (W) wide, and negative (grey) when the reverse was true.

Table 1. Substitution rate contrasts between woody (W) and herbaceous (H) sister pairs. A random subset of 3000 extracted Caryophyllales clades were used for each sister pair, except for contrast E, in which we analyzed all clades. Contrast values either lower than -10 or higher than 10 were excluded. *P* values were calculated from sign tests, assuming the average substitution rates were equal between the woody and the herbaceous side at each sister pair. (A) Without filtering by number of tips and (B) values calculated from cases where the woody lineage and the herbaceous lineage had the same number of tips.

A

	Synonymous rates				Nonsynonomous rates			
	# W>H	# W<H	*P*(W=H)	Median contrast	# W>H	# W<H	*P*(W=H)	Median contrast
Contrast A	49	2951	< 2.2e-16	2.227	68	2932	< 2.2e-16	2.018
Contrast B	31	2969	< 2.2e-16	1.624	150	2850	< 2.2e-16	1.382
Contrast C	1065	1917	< 2.2e-16	0.564	935	2008	< 2.2e-16	0.605
Contrast D	165	2835	< 2.2e-16	1.176	336	2664	< 2.2e-16	0.904
Contrast E	188	534	< 2.2e-16	0.366	231	491	< 2.2e-16	0.331

B

	Synonymous rates				Nonsynonomous rates			
	# W>H	# W<H	*P*(W=H)	Median contrast	# W>H	# W<H	*P*(W=H)	Median contrast
Contrast A	14	160	< 2.2e-16	1.950	22	152	< 2.2e-16	1.484
Contrast B	23	874	< 2.2e-16	1.434	93	804	< 2.2e-16	1.199
Contrast C	1065	1917	< 2.2e-16	0.564	935	2008	< 2.2e-16	0.605
Contrast D	66	180	2.1e-13	0.715	84	162	7.5e-7	0.461
Contrast E	188	534	< 2.2e-16	0.366	231	491	< 2.2e-16	0.331

Table 2. Caryophyllales clades with the highest number of tips. (A) Clades with the highest number of tips for any single taxon. (B) Clades with the highest total number of tips.

A

Clade ID	# tips	# taxa	Taxa with highest copies:#copies	Annotation
cc2163-1.mm.cary	105	30	Nyctaginaceae_Anulocaulis_leiosolenus_H:18	Probable nucleoredoxin 1-like [*Citrus sinensis*]
cc123-2.mm.1.cary	354	65	Amaranthaceae_Alternanthera_brasiliana_H:17; Caryophyllaceae_Spergularia_media_H:13	Oxidoreductase family protein [*Populus trichocarpa*]
cc42-2.mm.2.cary	74	23	Nyctaginaceae_Allionia_incarnata2_H:16; Nyctaginaceae_Allionia_incarnata_H:16	Retrotransposon protein, putative, Ty1-copia subclass [*Oryza sativa*, Japonica Group]
cc504-1.mm.1.cary	184	41	Aizoaceae_Sesuvium_ventricosum_H:15; Nyctaginaceae_Allionia_incarnata2_H:13	Mariner transposase [*Beta vulgaris* subsp. *vulgaris*]
cc3-6.mm.1.cary	218	54	Portulacaceae_Portulaca_molokiniensis_H:15	Contains similarity to reverse transcriptases [*Arabidopsis thaliana*]
cc327-1.mm.3.cary	56	18	Amaranthaceae_Atriplex_rosea_H:15	Protein FAR1-related sequence 5-like [*Citrus sinensis*]

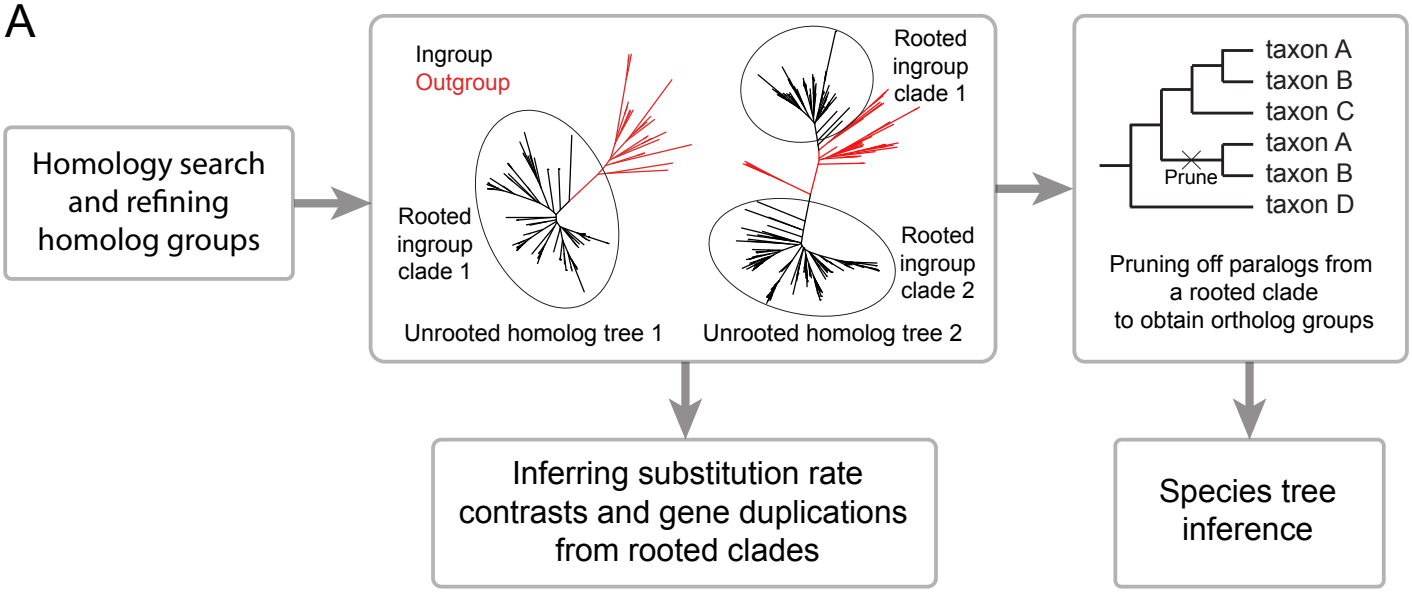
cc1593-1.mm.1.cary	56	18	Nyctaginaceae_Boerhavia_burbidgeana_H:14	Hypothetical protein VITISV_026753 [*Vitis vinifera*]
cc3-8.mm.cary	56	18	Nyctaginaceae_Boerhavia_burbidgeana_H:14	Integrase core domain containing protein [*Solanum demissum*]
cc28-1.mm.1.cary	353	68	Nyctaginaceae_Boerhavia_coccinea_H:13	Cytochrome P450 71A25-like [*Fragaria vesca* subsp. *vesca*]
cc6-7.mm.1.cary	99	30	Nyctaginaceae_Boerhavia_coccinea_H:13	Putative reverse transcriptase [*Arabidopsis thaliana*]
cc27-3.mm.1.cary	279	62	Amaranthaceae_Aerva_lanata_H:13	G-type lectin S-receptor-like serine/threonine-protein kinase At1g11330-like [*Vitis vinifera*]; recognition of pollen
cc1593-1.mm.1.cary	97	18	Nyctaginaceae_Allionia_incarnata2_H:13	Uncharacterized protein LOC100259102 [*Vitis vinifera*]
cc5-17.mm.1.cary	352	59	Nyctaginaceae_Guapira_obtusata_W:13; Amaranthaceae_Atriplex_hortensis_H:13	NBS-LRR type resistance protein [*Beta vulgaris*]

B

Clade ID	# tips	# taxa	Annotation
cc25-1.mm.1.cary	361	68	Cytochrome P450, family 72, subfamily A, polypeptide 15 isoform 2 [*Theobroma cacao*]
cc123-2.mm.1.cary	354	65	Oxidoreductase family protein [*Populus trichocarpa*]
cc28-1.mm.1.cary	353	68	Cytochrome P450 71A25-like [*Fragaria vesca* subsp. *vesca*]
cc5-17.mm.1.cary	352	59	NBS-LRR type resistance protein [*Beta vulgaris*]
cc333-1.mm.1.cary	302	63	Cytochrome P450 76C1-like [*Vitis vinifera*]
cc144-1.mm.1.cary	287	68	Peptide transporter PTR2-like [*Solanum tuberosum*]
cc171-1.mm.1.cary	285	61	Zeatin O-glucosyltransferase-like [*Fragaria vesca* subsp. *vesca*]
cc27-3.mm.1.cary	279	62	G-type lectin S-receptor-like serine/threonine-protein kinase At1g11330-like [*Vitis vinifera*]
cc521-1.mm.cary	277	67	GDSL esterase/lipase 1-like [*Vitis vinifera*]
cc50-2.mm.1.cary	273	63	Wall-associated receptor kinase-like 9-like [*Vitis vinifera*]

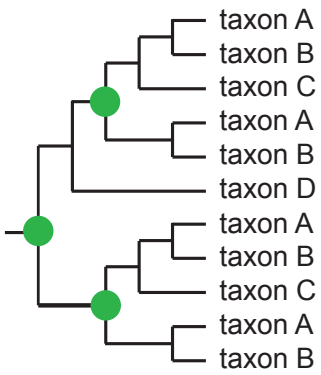
Figure 1

A



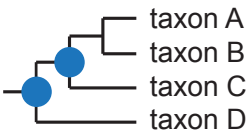
B

● = Locations of gene duplication detected from a homolog tree



Rooted clades extracted from homologs

● = Gene duplication events from a homolog tree mapped to the species tree. When nested duplications are detected, only one event per node is counted



Species tree

Figure 2

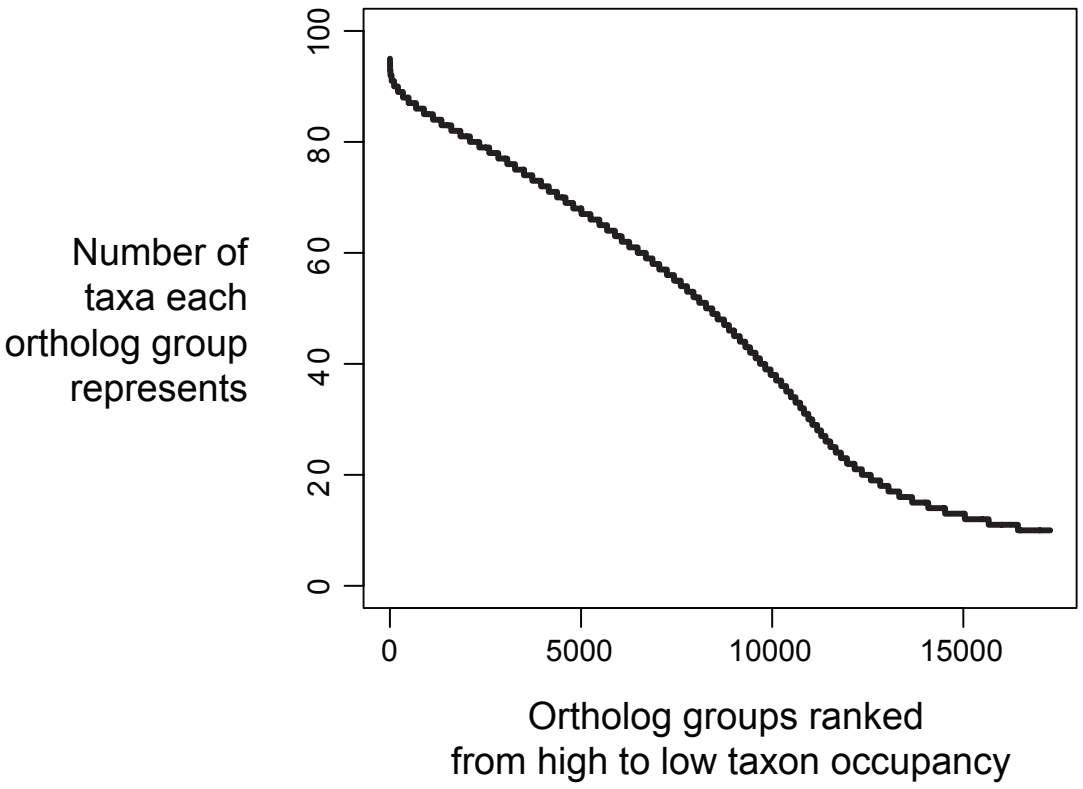


Figure 3

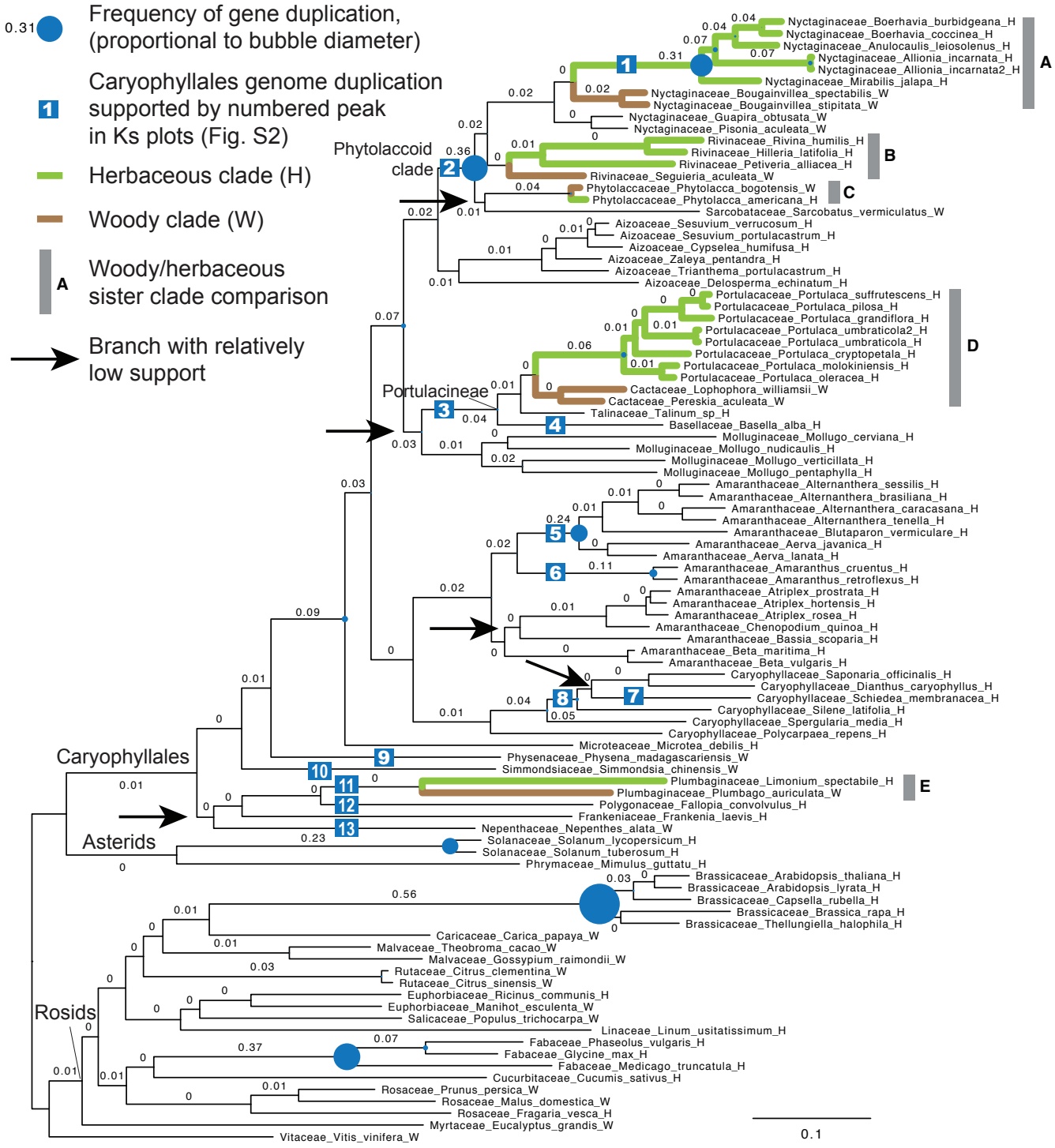


Figure 4