Discovery Article

Deep phylogenetic incongruence in the angiosperm clade *Rosidae*

MIAO SUN[1, 2], DOUGLAS E. SOLTIS[*3, 4, 5], PAMELA S. SOLTIS[4, 5], XINYU ZHU[6], J. GORDON BURLEIGH[3, 5], ZHIDUAN CHEN*[1]

[1]*State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, the Chinese Academy of Sciences, Beijing 100093, China;*

[2]*Graduate University of the Chinese Academy of Sciences, Beijing 100039, China;*

[3]*Department of Biology, University of Florida, Gainesville, FL 32611, USA;*

[4]*Florida Museum of Natural History, University of Florida, Gainesville, FL 32611, USA;*

[5]*University of Florida Genetics Institute*

[6]*School of Life Science, Nantong University, Nantong 226007, China;*

***Corresponding authors**:

Zhiduan Chen

State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, the Chinese Academy of Sciences, 20 Nanxincun, Xiangshan, Beijing, 100093, China; zhiduan@ibcas.ac.cn.

Douglas Soltis

Florida Museum of Natural History, University of Florida, Gainesville, FL 32611, USA; dsoltis@ufl.edu.

ABSTRACT

Analysis of large data sets can help resolve difficult nodes in the tree of life and reveal complex evolutionary histories, including instances of lateral gene transfer, hybridization, or incomplete lineage sorting. The placement of the Celastrales-Oxalidales-Malpighiales (COM) clade within the large *Rosidae* clade remains one of the most difficult deep-level phylogenetic questions in angiosperms, with previous analyses placing it with either *Fabidae* or *Malvidae*. To elucidate the underlying cause of this phylogenetic discordance, we assembled taxonomically comparable multi-gene matrices of chloroplast, mitochondrial, and nuclear sequences, as well as large single- and multi-copy nuclear gene data sets. Analyses of multi-gene data sets demonstrate incongruence between the chloroplast and both nuclear and mitochondrial data sets, and the results are robust to various character-coding and data-exclusion treatments. Analyses of single- and multi-copy nuclear genes indicate that most loci support the placement of COM with *Malvidae*, with a notable number of genes supporting COM with *Fabidae*, and almost no support for COM outside a clade of *Malvidae* and *Fabidae*. The proportion of genes supporting each hypothesis suggests that the phylogenetic incongruence is not due to incomplete lineage sorting. Although ancient introgressive hybridization remains a plausible explanation for the conflict among genes, greater taxon sampling is necessary to evaluate this hypothesis fully. The results emphasize the importance of genomic data sets for revealing deep incongruence, and potentially complex patterns of evolution, in organismal phylogeny.

KEYWORDS

Hybridization, introgression, incomplete lineage sorting, COM clade, incongruence, phylogenomics

INTRODUCTION

Genome-scale data provide the power to resolve some of the most perplexing parts of the tree of life (e.g., Dunn et al. 2008; Lee et al. 2011; Smith et al. 2011; Simon et al. 2012; Yoder et al. 2013). However, estimates from numerous independent loci in also can reveal phylogenetic incongruence caused by different evolutionary processes, such as gene duplication and loss, recombination, hybridization, lateral gene transfer, or incomplete lineage sorting (e.g., Goodman et al. 1979; Hudson 1983; Doyle 1992; Maddison 1997; Degnan and Rosenberg 2009; Cui et al. 2013; Oliver 2013).Molecular phylogenetic analyses have resolved much of the backbone angiosperm phylogeny (e.g., Soltis et al. 2009, 2011; Ruhfel et al. 2014) and clarified long-standing questions regarding relationships within major clades such as monocots (*Monocotyledoneae*; Chase et al. 2000; Jerrold et al. 2004; Graham et al. 2006; Givnish et al. 2006, 2010; Saarela et al. 2008), asterids (*Asteridae*; Olmstead et al. 2000; Albach et al. 2001; Bremer et al. 2001, 2004; Hilu et al. 2003; Moore et al. 2011), and rosids *(Rosidae;* Hilu et al. 2003; Jansen et al. 2007; Soltis et al. 2007, 2011; Wang et al. 2009; Moore et al. 2010; Qiu et al. 2010), yet much of this work is based either largely or exclusively on chloroplast sequence data, which represent a single, linked locus. With next-generation sequencing technologies, nuclear markers are now available to evaluate results from analyses of chloroplast gene sequence data and potentially reveal phylogenetic incongruence among loci (e.g., Duarte et al. 2010; Burleigh et al. 2011; Lee et al. 2011).

Introgressive hybridization has played an important role in plant evolution, and incomplete lineage sorting also is likely to have occurred during some rapid radiations. Consequently, there are numerous examples of discordance between chloroplast and nuclear gene trees in plants (e.g., Rieseberg and Soltis 1991; Rieseberg and Wendel 1993; Soltis and Kuzoff 1995; Wendel et al. 1995; Rieseberg et al. 1995, 1996a; Tsitrone et al. 2003; Okuyama et al. 2005; Soltis and Soltis 2009; Acosta and Premoli 2010). Although phylogenetic analyses based on nuclear, mitochondrial, and chloroplast loci across most of the angiosperm backbone tree have largely agreed, one major point of discordance in the large

Rosidae clade is the placement of the COM (Celastrales-Oxalidales-Malpighiales; Endress and Matthews 2006; Zhu et al. 2007) clade.

Rosidae comprises approximately one quarter of all angiosperm species and includes many temperate and tropical forest trees and shrubs. Due to rapid radiations (Wang et al. 2009), resolving relationships within *Rosidae* has been difficult (e.g., Hilu et al. 2003; Soltis et al. 2005, 2007, 2011; Jansen et al. 2007; Zhu et al. 2007; Wang et al. 2009; Moore et al. 2010, 2011; Qiu et al. 2010; Lee et al. 2011; Ruhfel et al. 2014). However, multi-gene studies have recovered two major, well-supported clades — the *Fabidae* (i.e., eurosids I, fabids) and *Malvidae* (i.e., eurosids II, malvids) (Soltis et al. 1999, 2000, 2005, 2007, 2011; Hilu et al. 2003; Judd and Olmstead 2004; Wang et al. 2009; Moore et al. 2010, 2011).

The COM clade contains approximately one third of all *Rosidae*, 870 genera and ~19,000 species (APG III 2009). Molecular analyses, largely dominated by chloroplast genes, generally have placed the COM clade with *Fabidae* (Table 1; e.g., Hilu et al. 2003; Soltis et al. 2005, 2007, 2011; Jansen et al. 2007; Burleigh et al. 2009; Wang et al. 2009; Moore et al. 2010, 2011). Analyses of the mitochondrial gene *matR* first suggested the placement of the COM clade with *Malvidae* (Zhu et al. 2007), and subsequent studies based on nuclear or mitochondrial genes supported this placement, although typically with limited taxon sampling (Table 1; Duarte et al. 2010; Finet et al. 2010; Qiu et al. 2010; Shulaev et al. 2010; Burleigh et al. 2011; Lee et al. 2011; Morton 2011; Zhang et al. 2012). Several floral characters also appear to link the COM clade with *Malvidae*. For example, in COM and *Malvidae* species the inner integument of the ovule is thicker than the outer integument at the time of fertilization, a feature that is extremely rare in *Fabidae* and other eudicots. Additionally, contorted petals and a tendency towards polystemony and polycarpy also suggest a placement of COM members with *Malvidae* rather than with *Fabidae* (Endress and Matthews 2006; Endress et al. 2013).

Although analyses of chloroplast gene sequence data generally appear to conflict with analyses of mitochondrial and nuclear gene sequence data, studies conducted to date differ greatly in taxon sampling and analytical methods (Table 1). Thus, it is unclear whether the

different placements in the COM clade are due to error in the analyses or biological incongruence. The level of incongruence within the nuclear genome also is unknown. We use the COM clade as an exemplar to investigate phylogenetic incongruence at deep levels in angiosperm phylogeny. Specifically, we first compare phylogenetic results from chloroplast, mitochondrial, and nuclear data sets having similar taxon sampling and determine whether the results are robust to various character-coding and data-exclusion protocols. We also survey large-scale nuclear data sets of both single-copy and multi-copy genes to investigate the patterns of phylogenetic discordance within the nuclear genome and then discuss whether these patterns are consistent with incomplete lineage sorting (deep coalescence) (Maddison 1997; Page and Charleston 1998; Maddison and Knowles 2006) or ancient hybridization and introgression (Tsitrone et al. 2003; Linder and Rieseberg 2004; Chang et al. 2011; Cui et al. 2013).

RESULTS

Chloroplast, Mitochondrial, and Nuclear Data Sets

Maximum likelihood (ML) analyses of the chloroplast, mitochondrial, and nuclear multi-gene alignments with similar taxon sampling recover conflicting placements of the COM clade (Figure 1–3). We focus on the relationships among members of *Rosidae*, but the full trees for all analyses are available as supplemental data and on Dryad (XXXX).

The phylogeny based on the 82-taxon, 78-gene chloroplast data set largely agrees with conclusions from previous chloroplast-dominated studies (APG III 2009; Wang et al. 2009; Moore et al. 2010; Soltis et al. 2011; Ruhfel et al. 2014), supporting the placement of the COM clade with *Fabidae* (Figure 1). The COM clade received 100% BS support, as did a clade of all COM and *Fabidae* species, but the precise placement of the COM clade was uncertain. There was 52% BS support for a sister relationship of the COM clade and all *Fabidae* except *Bulnesia* (Figure 1), which was sister to COM and other *Fabidae* species. Although most chloroplast genes support a COM with *Fabidae*, albeit generally with low BS support, no chloroplast genes have even 10% BS support for COM with *Malvidae* (see supplementary file S4). The analysis of the full chloroplast alignment with RY coding

indicates 100% BS support for a clade of COM and *Fabidae* species (supplementary file S4). AA coding indicates 47% BS support for a clade of the COM species and all *Fabidae* except *Bulnesia*, which is placed in *Malvidae*, although with low support (supplementary file S4). Removing the highly variable nucleotide sites from the chloroplast alignment quickly erodes support for COM with *Fabidae* (supplementary file S4). Bootstrap support for COM with *Fabidae* drops from 98% to 48% to 1% with the removal of the 5%, 10%, and 20% most variable sites, respectively, with no support after removing more sites. However, none of the site removal analyses indicates any support for a clade of COM and *Malvidae* or COM outside of *Fabidae* + *Malvidae* (supplementary file S4).

Trees from analyses of the 79-taxon, 4-gene mitochondrial data set generally indicate a close relationship of the COM clade with *Malvidae* (Figure 2; supplementary file S5). In the ML analysis of the full nucleotide data set, there is 94% BS support for a clade of COM clade species and all *Malvidae* except *Stachyurus* and *Oenothera* (Figure 2). Additionally, *Guaiacum* (Zygophyllaceae, Zygophyllales) is sister to *Stachyurus* (Stachyuraceae, Crossosomatales) in agreement with Qiu et al. (2010). However, some relationships obtained here are inconsistent with previous analyses. For example, the placements of *Garrya* and *Guaiacum* differ compared to those obtained in studies based largely on chloroplast or nuclear genes (supplementary file S5; see APG III 2009; Soltis et al. 2011). Analyses of the four individual mitochondrial genes show either weak (< 60%) BS support linking COM with *Malvidae* or are unresolved, with little, if any, support for either clade or *Rosidae* (supplementary file S5). AA coding indicates 74% BS support for a clade of COM species and all *Malvidae* except *Stachyurus* and *Oenothera* (supplementary file S5). RY coding greatly reduces support for *Rosidae* relationships, with no support even for the COM clade. Removing the most variable 5% of sites yields 100% BS support for a clade of COM species and all *Malvidae* except *Stachyurus* and *Oenothera.* However, removing more variable sites greatly reduces support for relationships throughout the tree; after removing the 10% most variable sites, BS support for the COM clade drops to 23% (supplementary file S5).

The results from analyses of the 92-taxon, 5-gene nuclear data set provide 100% BS support for a clade that includes COM species and all species of *Malvidae* except *Pelargonium*, *Oenothera*, and *Lagerstroemia* (Figure 3), as do the results from ML analyses of the RY and AA matrices (supplementary file S6). This placement of COM with *Malvidae* is also in agreement with the ML analyses of the five individual nuclear genes, although to various degrees (supplementary file S6). Likewise, the ML analyses of nucleotides after removing 5% and 10% of the most variable sites yield 100% BS support for a clade that includes all of the COM species and *Malvidae* species, except *Pelargonium*, *Oenothera*, and *Lagerstroemia* (supplementary file S6). Removing 20% or 30% of the most variable sites reduces the support for this clade to 96% and 94%, respectively, but removing more sites greatly reduces support for relationships within *Rosidae* in general, including support for the monophyly of the COM clade (supplementary file S6).

Single-copy Nuclear Gene Analysis

Although most of the orthologous gene sets from the Lee et al. (2011) analysis were not informative regarding the placement of the COM clade, of those genes that do support one of the three possible placements (COM with *Fabidae*, COM with *Malvidae*, or COM outside of *Fabidae* + *Malvidae*), 61–75% support a clade of COM with *Malvidae* (Table 2), 25-39% support a clade of COM with *Fabidae,* and none of the orthologous gene sets support COM outside of *Fabidae* + *Malvidae* (Table 2). While increasing the minimum bootstrap support cutoff reduces the number of orthologous gene sets supporting the hypotheses, it has little effect on the relative number of genes supporting COM with *Malvidae* versus COM with *Fabidae* (Table 2).

Multi-copy Nuclear Gene Analysis

Similar to the single-copy nuclear gene analysis, most of the multi-copy gene trees were not informative regarding the placement of the COM clade, with the majority of informative genes supporting the placement of COM with *Malvidae*. Between 71–98% of the informative genes support a clade of COM and *Malvidae* species under each of the different reconciliation costs (Table 3). The duplication-only model provides the strongest support for the COM with

Malvidae clade (≥ 91%; Table 3). The maximum percentage of informative genes supporting COM with *Fabidae* is 27%, based on the deep coalescence reconciliation model (Table 3). Support for a clade of COM outside of *Fabidae* + *Malvidae* ranges from 0–6% of the genes in these analyses (Table 3).

DISCUSSION

In spite of much recent progress resolving the angiosperm tree of life, the phylogenetic placement of the COM clade remains uncertain. Previous efforts to place the COM clade have used a variety of data sources, taxon sampling strategies, and phylogenetic methods. Therefore, it is difficult to determine if the conflicting placements of the COM clade are due to error or actual biological conflict among loci (Table 1). Our ML analyses of multi-gene chloroplast, mitochondrial, and nuclear data sets with similar taxon sampling reinforce the observation that analyses of chloroplast loci yield a topology that differs from analyses of mitochondrial and most nuclear loci (Table 1; Figure 1–3). These analyses are robust to different character-coding strategies, which are often used to detect heterogeneous phylogenetic signals or error. AA matrices and RY coding are used to ameliorate nucleotide saturation and composition biases (Hashimoto et al. 1995; Phillips and Penny 2003; Harrison et al. 2004; Delsuc et al. 2005; Gibson et al. 2005), and removal of highly variable sites has been proposed to reduce long-branch attraction or model-fitting error (see Philippe et al. 2005). Some of these experiments erode phylogenetic signal for the placement of the COM clade, but none support an alternate placement of the COM clade. Although we failed to find obvious signs of major systematic or sampling biases, it is difficult to demonstrate the absence of error. In fact, the (weakly supported) variation in single-gene topologies of linked chloroplast genes suggests that some level of error may be present in chloroplast gene sequence analyses (supplementary file S4). Nonetheless, the consistency of the incongruence suggests that there may be an underlying biological basis to the conflict among chloroplast, nuclear, and mitochondrial loci.

If the conflict among chloroplast, nuclear, and mitochondrial gene sequence data is due to evolutionary events such as ancient hybridization or incomplete lineage sorting, we would

also expect to see conflict among independent nuclear loci. Indeed, within the single-copy nuclear gene data set from Lee et al. (2011), 61–75% of the informative genes support a placement of COM with *Malvidae*, while 25–39% support the placement of COM with *Fabidae* (Table 2). The multi-copy genes reveal similar levels of incongruence, with ≥ 71% of the informative genes supporting COM with *Malvidae,* with far less support for COM with *Fabidae* and very little support for COM outside a clade of *Malvidae* + *Fabidae* (Table 3). This predominant placement of COM with *Malvidae* is consistent with large gene tree parsimony (Burleigh et al. 2011; Górecki et al. 2012). The placement of the COM clade from multi-copy genes is robust to the model of gene reconciliation (Table 3). Furthermore, in both the single- and multi-copy gene results, the overall percentage of informative genes supporting each of the three hypotheses is relatively stable no matter the bootstrap cutoff we use (Table 2, 3).

If the differences in the position of the COM clade among nuclear loci are not due to error, they may reflect ancient hybridization and/or incomplete lineage sorting. Distinguishing between incomplete lineage sorting and hybridization can be challenging (e.g., Sang and Zhong 2000; Holder et al. 2001; Buckley et al. 2006; Holland et al. 2008; Joly et al. 2009), and the sparse and incomplete taxon sampling within our nuclear gene data sets, as well as the ancient divergence time of the major rosid lineages (Bell et al. 2010; Wang et al. 2009), make it especially difficult to differentiate between the two. Although the effects of incomplete lineage sorting typically are studied on recent radiations, incomplete lineage sorting can also affect the resolution of ancient radiations (Whitfield and Lockhart 2007; Oliver 2013), such as the deep relationships among mammals (McCormack et al. 2012; Song et al. 2012). However, if we consider the placement of the COM clade as a rooted 3-taxon (COM, *Fabidae*, and *Malvidae*) phylogenetic problem, a process of incomplete lineage sorting should yield approximately equal numbers of nuclear genes supporting the two possible non-species tree topologies (Huson et al. 2005). Instead, we see that the majority of genes supports COM with *Malvidae* and, to a lesser extent, COM with *Fabidae*, with almost no support for COM outside of *Fabidae* + *Malvidae* (Tables 2, 3). This pattern of support for only two of the three

possible 3-taxon topologies suggests that incomplete lineage sorting does not explain the phylogenetic discordance among genes. This result is consistent with recent model-fitting analyses indicating that the multispecies coalescent is a poor fit for many phylogenetic data sets (Reid et al. 2013).

Many plant lineages have experienced hybridization and introgression throughout their evolutionary histories (e.g., Okuyama et al. 2005), and there are more than a hundred records of interspecific hybridization among rosid taxa alone (Rieseberg and Soltis 1991; Rieseberg et al. 1996a). An ancient introgressive hybridization event likely would produce discordance among independent loci. The incongruent placement of the COM clade in trees constructed from mitochondrial and chloroplast gene sequence data suggests that the evolutionary histories of these two subcellular compartments are unlinked, with the chloroplast genome derived from the *Fabidae* lineage and the mitochondrial genome from the *Malvidae* lineage. This result is unexpected given that the chloroplast and mitochondrial genomes typically are both maternally inherited in angiosperms (Birky 1995, 2001; Corriveau and Coleman 1988; Mogensen 1996). However, there are cases of biparental inheritance of organellar genomes (e.g., Fauré et al. 1994; Testolin and Cipriani, 1997; Havey et al. 1998; Yang et al. 2000), and paternal inheritance of chloroplast genomes has been documented in species in the COM clade (*Turnera ulmifolia*; Shore et al. 1994; Shore and Triassi 1998) and *Fabidae* (*Medicago sativa*; Schumann and Hancock 1989; Masoud et al. 1990; *Larrea*; Yang et al. 2000). Also, empirical studies suggest that progeny from a hybridization event may exhibit strong paternal chloroplast inheritance, while mitochondrial inheritance remained exclusively maternal (Schumann and Hancock 1989; Masoud et al. 1990; Shore et al. 1994; Xu 2005). Thus, it is conceivable that an ancient hybridization event resulted in different evolutionary histories for the chloroplast and mitochondrial genomes.

In this putative ancient hybridization scenario, an early member of *Fabidae* or its immediate ancestor acted as the paternal parent and crossed with the maternal lineage of a member of *Malvidae,* with accompanying chloroplast paternal transmission to the ancestor of the COM clade (F_1). This event could have created conflicting histories in the chloroplast and

mitochondrial genomes and conflict among nuclear loci with half of the alleles in the F_1 contributed by each parent (Figure 1, 2, 4). Repeated selfing or crossing of the hybrid derivatives would not explain the high percentage of nuclear loci supporting the relationship of COM with *Malvidae* (Tables 2, 3). Thus, there likely were subsequent backcrosses of the early hybrids to the maternal *Malvidae*, reducing the number of nuclear loci supporting the placement of COM with *Fabidae* (Figure 4).

Numerous plant systematics studies have demonstrated the promise of genomic data to resolve angiosperm relationships that were not evident in analyses with a few genes (Finet et al. 2010; Moore et al. 2010, 2011; Burleigh et al. 2011; Lee et al. 2011). We demonstrate here that analyses of data sets with many unlinked loci can highlight the ambiguity and discordance in phylogenetic relationships and potentially reveal the complexity of angiosperm evolution. Most, but not all, single and multi-copy nuclear loci, as well as mitochondrial genes, support the placement of the COM clade with *Malvidae*. This placement is also consistent with patterns of morphological evolution (Endress and Matthews 2006), but it contradicts the strongly supported analyses of chloroplast sequence datasets (Figure 1–4; Jansen et al. 2007; Moore et al. 2010, 2011; Ruhfel et al. 2014). While analyses involving a single data source, such as the chloroplast genome, seek a single phylogeny, it may be more informative to appreciate the potentially chimeric origins of the COM clade rather than to force its placement in a binary species tree. Although evidence of ancient reticulate evolution near the origin of the COM clade is not conclusive with current sampling, our analyses emphasize the importance of phylogenomic data for highlighting phylogenetic incongruence and directing future studies.

METHODS

Throughout the paper, to facilitate discussion, we treat COM, *Fabidae*, and *Malvidae* as three separate groups, despite current classifications that consider the COM clade to be part of *Fabidae* (Cantino et al. 2007) or fabids (APG III 2009).

Phylogenetic Analyses of Chloroplast, Mitochondrial, and Nuclear Data

To compare the placement of the COM clade in analyses of chloroplast, mitochondrial and nuclear gene data sets, we assembled published matrices with similar taxon sampling. For the chloroplast gene sequence data, we pruned 82 seed plant taxa from the 78-gene chloroplast data set of Ruhfel et al. (2014). We also used the 92-taxon, 5-gene nuclear data set of Zhang et al. (2012), and the 79-taxon, 4-gene mitochondrial matrix of Qiu et al. (2010). The taxon sampling in all of these studies was designed to reconstruct relationships across angiosperms using representative sampling of major clades, including the COM clade. Based on the nuclear gene sequence data set of Zhang et al. (2012), we assembled the chloroplast and mitochondrial gene data sets, attempting to ensure as much as possible that the taxa employed from these data sets are from the same species or genus. The familial and ordinal names of the sampled taxa follow APG III (2009).

We performed a series of maximum likelihood (ML) phylogenetic analyses on each of the three data sets using RAxML v.7.2.8 (Stamatakis 2008). For all analyses, we estimated the optimal ML tree and performed 100 nonparametric bootstrap (BS) replicates. First, we analyzed the full nucleotide alignments using an unpartitioned GTRCAT model. We also examined the variation of the COM clade placement in single gene topologies inferred from these three data sets using RAxML with the GTRCAT model. For the three multi-gene data sets, we also analyzed the amino acid (AA) alignment using the PROTCATJTT model (Jones et al. 1992). RY coding, which recodes the nucleotides as binary characters, either purines (A or G = R) or pyrimidines (C or T = Y), has been used to ameliorate biases caused by saturation, rate heterogeneity, and base composition (Phillips and Penny 2003; Harrison et al. 2004; Phillips et al. 2004; Delsuc et al. 2005; Gibson et al. 2005). Thus, we also transformed the three full nucleotide matrices to RY coding and ran a ML analysis using the GTRCAT model.

The elimination of potentially misleading sites from an alignment is a common practice in phylogenetic analysis (e.g. Delsuc et al. 2005; Philippe et al. 2005; Regier and Zwick 2011; Rajan 2013). We removed highly variable sites as a further means of exploring the data that may contribute to the discordant placements of the COM clade. Following the method

described by Goremykin et al. (2010), we organized the nucleotide sites in each alignment in order of rate based on the observed variability (OV) criterion. For each sorted alignment, we then removed the most variable 5%, 10%, 20%, 30%, 40%, and 50% of the sites, and after each removal performed an ML analysis on the remaining sites in each alignment using the GTRCAT model.

Single-Copy Nuclear Gene Analysis

The largest nuclear gene data set used to resolve the backbone of angiosperm relationships comprises 22,833 groups of orthologs (Lee et al. 2011). Although this data set includes only seven species of Malpighiales representing the COM clade (no Celastrales or Oxalidales species were included), it provides estimates from by far the greatest number of presumably independent nuclear loci for the placement of the COM clade. We examined the individual gene trees from this data set to look for variation in the placement of the COM clade. First, we divided the full, concatenated nucleotide alignment from Lee et al. (2011; available on the BIGPLANT website: http://nybg.bio.nyu.edu/) into separate alignments (see supplementary file S1), each representing a set of putative orthologs. Next we identified the ortholog sets that were potentially informative regarding the placement of the COM clade; these alignments contained at least one COM species, one *Malvidae*, one *Fabidae*, and one other species not in any of these groups. This criterion resulted in 8,445 potentially informative ortholog sets. For each ortholog set alignment, we performed 100 ML bootstrap replicates using RAxML v.7.2.8 with the GTRCAT model (Stamatakis 2008), and we counted how many bootstrap replicates support a clade of COM and *Fabidae* species, how many support a clade of COM and *Malvidae* species, and how many support COM outside a clade of *Fabidae* and *Malvidae*. The analysis of the support for the COM placement was automated using Perl scripts and Newick utilities (Junier and Zdobnov 2010).

Multi-copy Nuclear Gene Analysis

We also examine support for the placement of the COM clade using multi-copy nuclear gene families, or gene families that may have multiple sequences from one or more taxa. Unlike single-copy sets of orthologs, it is not always straightforward to interpret the

phylogenetic signal supported by multi-copy gene families. For example, a species from the COM clade could have one sequence that groups with *Fabidae* and one sequence that groups with *Malvidae*. To solve this problem, we estimated the reconciliation cost of each gene family tree, given a topology with COM sister to *Fabidae*, COM sister to *Malvidae*, and COM sister to a *Fabidae* + *Malvidae* clade based on three different reconciliation costs, each implying a different evolutionary scenario: 1) the minimum number of implied gene duplications; 2) the minimum number of implied duplications and losses; and 3) the minimum number of implied deep coalescence events (e.g., Maddison 1997). We used a parsimony criterion to distinguish among the three species tree topologies; the topology with the lowest reconciliation cost, or the topology that implies the fewest evolutionary events, is the topology that is supported by the gene family. If two or three of the topologies have equal reconciliation costs, the gene family is considered uninformative regarding the placement of COM.

We assembled a collection of 3,748 gene family alignments (see supplementary file S2) obtained from the genome sequences of 22 plant taxa with OrthoMCL (Chen et al. 2006) and aligned with MAFFT (Katoh et al. 2005). Included are *Selaginella moellendorffii*, *Physcomitrella patens,* and 20 angiosperm species, including one species representing the COM clade, *Populus trichocarpa* (supplementary file S3). Although the taxon sampling is sparse, using only sequences from completely sequenced genomes may enable more accurate estimates of processes such as gene loss than incomplete transcriptome data sets.

For each of the multi-copy gene alignments, we performed 100 bootstrap replicates using RAxML v.7.2.8 with the GTRCAT model (Stamatakis 2008). For each of the resulting bootstrap trees, we calculated the reconciliation cost under the three different cost models (duplications, duplications and losses, and deep coalescence) using a species tree in which *Populus* (COM clade) was sister to *Fabidae* (*Fragaria vesca*, *Medicago trunculata*, and *Glycine max*), one in which *Populus* was sister to *Malvidae* (*Arabidopsis thaliana*, *Thellungiella parvula*, *Carica papaya*, and *Theobroma cacao*), and one in which *Populus* was sister to *Fabidae* + *Malvidae*. The rooting of gene trees can greatly affect the estimates of the

reconciliation cost, and it is often difficult to know how to root a multi-copy gene tree. Thus, for each gene tree, we used a rooting that minimized the reconciliation cost. We calculated the reconciliation costs for each gene tree bootstrap replicate under both species trees using the program OptRoot, written by Andre Wehe and available at http://www.wehe.us/optroot.html.

All data sets and results are available on Dryad (XXX; www.datadryad.org).

ACKNOWLEDGMENTS

We thank Yin-Long Qiu, who contributed to the early design of this project, and Ning Zhang, who graciously provided us with the 92-taxon, 5-gene nuDNA alignment used in this study.

This work was supported by the National Natural Science Foundation of China (NNSF 31270268), National Basic Research Program of China (No. 2014CB954101), Chinese Academy of Sciences Visiting Professorship for Senior International Scientists (grant number 2011T1S24), and the US National Science Foundation (DEB-1301828).

REFERENCES

Acosta MC, Premoli AC. 2010. Evidence of chloroplast capture in South American Nothofagus (subgenus *Nothofagus*, Nothofagaceae). Mol Phylogenet Evol. 54:235-242.

Albach DC, Soltis PS, Soltis DE, Olmstead RG. 2001. Phylogenetic analysis of asterids based on sequences of four genes. Ann Mo Bot Gard. 88:163-212.

APG III. 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. Bot J Linn Soc. 161:105-121.

Bell CD, Soltis DE, Soltis, PS. 2010. The age and diversification of the angiosperms re-revisited. Am J Bot. 97:1296-1303.

Birky CW. 1995. Uniparental inheritance of mitochondrial and chloroplast genes: mechanisms and Evolution. Proc Natl Acad Sci USA. 92:11331-11338.

Birky CW. 2001. The inheritance of genes in mitochondria and chloroplasts: laws, mechanisms, and models. Annu Rev Genet. 35:125-148.

Bremer K, Backlund A, Sennblad B, Swenson U, Andreasen K, Hjertson M, Lundberg J, Backlund M, Bremer B. 2001. A phylogenetic analysis of 100+ genera and 50+ families of euasterids based on morphological and molecular data with notes on possible higher level morphological synapomorphies. Pl Syst Evol. 229:137-169.

Bremer K, Friis E, Bremer B. 2004. Molecular phylogenetic dating of asterid flowering plants shows early Cretaceous diversification. Syst Biol. 53:496-505.

Buckley TR, Cordeiro M, Marshall DC, Simon C. 2006. Differentiating between hypotheses of lineage sorting and introgression in New Zealand alpine cicadas (*Maoricicada Dugdale*). Syst Biol. 55:411-425.

Burleigh JG, Hilu KW, Soltis DE. 2009. Inferring phylogenies with incomplete data sets: a 5-gene, 567-taxon analysis of angiosperms. BMC Evol Biol. 17:61.

Burleigh JG, Bansal MS, Eulenstein O, Hartmann S, Wehe A, Vision TJ. 2011. Genome-Scale Phylogenetics: Inferring the Plant Tree of Life from 18,896 Gene Trees. Syst Biol. 60:117-25.

Cantino PD, Doyle JA, Graham SW, Judd WS, Olmstead RG, Soltis DE, Soltis PS, Donoghue MJ. 2007. Towards a phylogenetic nomenclature of Tracheophyta. Taxon. 56:822-846.

Chang SW, Oshida T, Endo H, Nguyen ST, Dang CN, Nguyen DX, Jiang X, Li ZJ, Lin LK. 2011. Ancient hybridization and underestimated species diversity in Asian striped squirrels (genus *Tamiops*): inference from paternal, maternal and biparental markers. J Zool. 285:128-138.

Chase MW, Soltis DE, Olmstead RG, Morgan D, Les DH, Mishler BD, Duvall MR, Price RA, Hills HG, Qiu YL, et al. 1993. Phylogenetics of seed plants: an analysis of nucleotide-sequences from the plastid gene *rbcL*. Ann Mo Bot Gard. 80:528-80.

Chase MW, Soltis DE, Soltis PS, Rudall PJ, Fay MF, Hahn WJ, Sullivan S, Joseph J, Molvray M, Kores PJ, et al. 2000. Higher-level systematics of the monocotyledons: An assessment of current knowledge and a new classification. In: Wilson KL, Morrison DA, editors. Monocots: Systematics and Evolution. Collingwood: CSIRO Publishing. p. 3-16.

Chen F, Mackey AJ, Stoeckert CJ Jr, Roos DS. 2006. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res. 34:D363-368.

Comes HP, Abbott RJ. 2001. Molecular phylogeography, reticulation, and lineage sorting in Mediterranean *Senecio* sect. *Senecio* (Asteraceae). Evolution. 55:1943-1962.

Corriveau JL, Coleman AW. 1988. Rapid screening method to detect potential biparental inheritance of plastid DNA and results over 200 angiosperm species. Am J Bot. 75:1443-1458.

Cui R, Schumer M, Kruesi K, Walter R, Andolfatto P, Rosenthal GG. 2013. Phylogenomics reveals extensive reticulate evolution in *Xiphophorus* fishes. Evolution. 67:2166-2179.

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol Evol. 24:332-340.

Delsuc F, Brinkmann FH, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet. 6:361-375.

Doyle JJ. 1992. Gene trees and species trees - molecular systematics as one-character taxonomy. Syst Bot. 17:144-163.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 7:214.

Duarte JM, Wall PK, Edger PP, Landherr LL, Ma H, Pires JC, Leebens-Mack J, dePamphilis CW. 2010. Identification of shared single copy nuclear genes in *Arabidopsis*, *Populus*, *Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. BMC Evol Biol. 10:61.

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. 452:745-749.

Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for ancient admixture between closely related populations. Mol Biol Evol. 28:2239-2252.

Endress PK, Matthews ML. 2006. Floral structure and systematics in four orders of rosids, including a broad survey of floral mucilage cells. Plant Syst Evol. 260:223-251.

Endress PK, Davis CC, Matthews ML. 2013. Advances in the floral structural characterization of the major subclades of Malpighiales, one of the largest orders of flowering plants. Ann Bot. 111:969-985.

Fauré S, Noyer JL, Carreel F, Horry JP, Bakry F, Lanaud C. 1994. Maternal inheritance of chloroplast genome and paternal inheritance of mitochondrial genome in bananas (*Musa acuminate*). Curr Genet. 25:265-269.

Finet C, Timme RE, Delwiche CF, Marlétaz F. 2010. Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. Curr Biol. 21:2217-2222.

Fontaine KM, Cooley JR, Simon C. 2007. Evidence for Paternal Leakage in Hybrid Periodical Cicadas (Hemiptera: *Magicicada spp*). PLoS ONE. 2:e892.

Graham SW, Zgurski JM, McPherson MA, Cherniawsky DM, Saarela JM, Horne ESC, Smith SY, Wong WA, O'Brien HE, Pires, JC, et al. 2006. Robust inference of monocot deep phylogeny using an expanded multigene plastid data set. Aliso. 22:3-20.

Gibson A, Gowri-Shankar V, Higgs PG, Rattray M. 2005. A comprehensive analysis of mammalian mitochondrial genome base composition and improved phylogenetic methods. Mol Biol Evol. 22:251-264.

Givnish TJ, Pires JC, Graham SW, McPherson MA, Prince LM, Patterson TB, Rai HS, Roalson ER, Evans TM, Hahn WJ, et al. 2006. Phylogeny of the monocotyledons based on the highly informative plastid gene *ndhF*: Evidence for widespread concerted convergence. In: Columbus JT, Friar EA, Porter JM, Prince LM, Simpson MG, editors. Monocots: Comparative Biology and Evolution Excluding Poales. California: Rancho Santa Ana Botanic Garden. p. 28-51.

Givnish TJ, Ames M, McNeal JR, dePamphilis CW, Graham SW, Pires JC, Stevenson DW, Zomlefer WB, Briggs BG, Duvall MR, et al. 2010. Assembling the tree of the monocotyledons: Plastome sequence phylogeny and evolution of Poales. Ann Mo Bot Gard. 97:584-616.

Goodman M, Czelusniak J, Moore GW, Romero-Herrera AE, Matsuda G. 1979. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed by globin sequences. Syst Zool. 28:132-163.

Górecki P, Burleigh JG, Eulenstein O. 2012. GTP supertrees from unrooted gene trees: linear time algorithms for NNI based local searches. In: Bioinformatics Research and Applications. Berlin: Springer. p. 102-114.

Goremykin VV, Nikiforova SV, Bininda-Emonds ORP. 2010. Automated removal of noisy data in phylogenomic analyses. J Mol Evol. 71:319-331.

Hahn MW. 2007. Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome Evolution. Genome Biol. 8:R141.

Harrison GA, McLenachan PA, Phillips MJ, Slack KE, Cooper A, Penny D. 2004. Four new avian mitochondrial genomes help get to basic evolutionary questions in the late Cretaceous. Mol Biol Evol. 21:974-983.

Hashimoto T, Nakamura Y, Kamaishi T, Nakamura F, Adachi J, Okamoto K, Hasegawa M. 1995. Phylogenetic place of mitochondrion-lacking protozoan, *Giardia lamblia*, inferred from amino acid sequences of elongation factor 2. Mol Biol Evol. 12:782-793.

Havey MJ, McCreight JD, Rhodes B, Taurick G. 1998. Differential transmission of the *Cucumis* organellar genomes. Theor Appl Genet. 97:122-128.

Hilu KW, Borsch T, Müller K, Soltis DE, Soltis PS, Savolainen V, Chase MW, Powell MP, Alice LA, Evans R, et al. 2003. Angiosperm Phylogeny Based on *matK* Sequence Information. Am J Bot. 90:1758-1776.

Holder MT, Anderson JA, Holloway AK. 2001. Difficulties in detecting hybridization. Syst Biol. 50:978-982.

Holland BR, Benthin S, Lockhart PJ, MoultonV, Huber KT. 2008. Using supernetworks to distinguish hybridization from incomplete lineage sorting. BMC Evol Biol. 8:202.

Hudson RR. 1983. Testing the constant-rate neutral allele model with protein sequence data. Evolution. 37:203-217.

Huson DH, Klöpper T, Lockhart PJ, Steel MA. 2005. Reconstruction of reticulate networks from gene trees. In: Proceedings of the Ninth International Conference on Research in Computational Molecular Biology. Heidelberg: Springer. p. 233-249.

Jansen RK, Cai Z, Raubeson LA, Daniell H, Depamphilis CW, Leebens-Mack MJ, Muller KF, Guisinger-Bellian M, Haberle RC, Hansen AK. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci USA. 104:19369-19374.

Jerrold I, Davis DW, Stevenson GP, Ole S, Lisa M, Campbell JV, Freudenstein DH, Goldman CR, Hardy FA, Michelangeli MP, et al. A phylogeny of the monocots, as inferred from *rbcL* and *atpA* sequence variation, and a comparison of methods for calculating jackknife and bootstrap values. Syst Bot. 29:467-510.

Joly S, McLenachan PA, Lockhart PJ. 2009. A statistical approach for distinguishing hybridization and incomplete lineage sorting. Am Nat. 174:E54-E70.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. CABIOS. 8:275-282.

Judd WS, Olmstead RG. 2004. A survey of tricolpate eudicot. phylogenetic relationships. Am J Bot. 91:1627-1644.

Junier T, Zdobnov EM. 2010. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. Bioinformatics. 26:1669-1670.

Katoh K, Kuma KI, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33:511-518.

Lee EK, Cibrian-Jaramillo A, Kolokotronis SO, 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 Genetics. 7:e1002411.

Linder C R, Rieseberg LH. 2004. Reconstructing patterns of reticulate evolution in plants. Am J Bot. 9:1700-1708.

Maddison WP. 1997. Gene trees in species trees. Syst Biol. 46:523-536.

Maddison WP, Knowles LL. 2006. Inferring phylogeny despite incomplete lineage sorting. Syst Biol. 55:21-30.

Masoud SA, Johnson LB, Sorensen EL. 1990. High transmission of paternal DNA in alfalfa plants demonstrated by restriction fragment polymorphic analysis. Theor Appl Genet. 79:49-55.

McCormack JE, Faircloth BC, Crawford NG, Gowaty PA, Brumfield RT, Glenn TC. 2012. Ultraconserved elements are novel phylogenetic markers that resolve placental mammal phylogeny when combined with species-tree analysis. Genome Res. 22:746-754.

Meng C, Kubatko LS. 2009. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. Theor Popul Biol. 75:35-45.

Mogensen HL. 1996. The hows and whys of cytoplasmic inheritance in seed plants. Am J Bot. 83:383-404.

Moore MJ, Soltis PS, Bell CD, Burleigh JG, Soltis DE. 2010. Phylogenetic analysis of 83 plastid genes further resolved the early diversification of eudicots. Proc Natl Acad Sci USA. 107:4623-4628.

Moore MJ, Hassan N, Gitzendanner MA, Bruenn RA, Croley M, Vandeventer A, Horn JW, Dhingra A, Brockington SF, Latvis M, et al. 2011. Phylogenetic analysis of the plastid inverted repeat for 244 species: insights into deeper-level angiosperm relationships from a long, slowly evolving sequence region. Int J Plant Sci. 172:541-558.

Morton MC. 2011. Newly Sequenced Nuclear Gene (*Xdh*) for Inferring Angiosperm Phylogeny. Ann Mo Bot Gard. 98:63-89.

Oliver JC. 2013. Microevolutionary processes generate phylogenomic discordance at ancient divergences. Evolution. DOI:101111/evo12047.

Olmstead RG, Kim KJ, Jansen RK, Wagstaff SJ. 2000. The phylogeny of the Asteridae sensu lato based on chloroplast *ndhF* gene sequences. Mol Phylogenet Evol. 16:96-112.

Okuyama Y, Fujii N, Wakabayashi M, Kawakita A, Ito M, Watanabe M, Murakami N, Kato M. 2005. Nonuniform concerted evolution and chloroplast capture: heterogeneity of observed introgression patterns in three molecular data partition phylogenies of Asian *Mitella* (Saxifragaceae). Mol Biol Evol. 22:285-296.

Page RDM, Charleston MA. 1998. Trees within trees: phylogeny and historical associations. Trends Ecol Evol. 13:356-359.

Phillips MJ, Penny D. 2003. The root of the mammalian tree inferred from whole mitochondrial genomes. Mol Phylogenet Evol. 28:171-185.

Phillips MJ, Delsuc F, Penny D. 2004. Genome-Scale Phylogeny and the Detection of Systematic Biases. Mol Biol Evol. 21:1455-1458.

Philippe H, Delsuc F, Brinkmann H, Lartillot N. 2005. Phylogenomics. Annu Rev Ecol Evol Syst. 36:541-562.

Qiu YL, Li LB, Wang B, Xue JY, Hendry TA, Li RQ, Brown JW, Liu Y, Hudson YH, Chen ZD. 2010. Angiosperm phylogeny inferred from sequences of four mitochondrial genes. J Syst Evol. 48:391-425.

Rajan V. 2013. A method of alignment masking for refining the phylogenetic signal of multiple sequence alignments. Mol Biol Evol. 30:689-712.

Rasmussen MD, Kellis M. 2011. A Bayesian approach for fast and accurate gene tree reconstruction. Mol Biol Evol. 28:273-290.

Rasmussen MD, Kellis M. 2012. Unified modeling of gene duplication, loss and coalescence using a locus tree. Genome Res. 22:755-65.

Regier JC, Zwick A. 2011. Sources of signal in 62 protein-coding nuclear genes for higher-level phylogenetics of arthropods. PLoS ONE. 6:e23408.

Reid NM, Hird SM, Brown JM, Pelletier TA, McVay JD, Satler JD, Carstens BC. 2013. Poor fit to the multispecies coalescent is widely detectable in empirical data. Syst Biol. syt057.

Rieseberg LH, Soltis DE. 1991. Phylogenetic consequences of cytoplasmic gene flow in plants. Evol Trends Plants. 5:65-84.

Rieseberg LH, Desrochers AM, Youn SJ. 1995. Interspecific pollen competition as a reproductive barrier between sympatric species of *Helianthus* (Asteraceae). Am J Bot. 82:515-519.

Rieseberg LH, Whitton J, Linder CR. 1996a. Molecular marker incongruence in plant hybrid zones and phylogenetic trees. Acta Bot Neerl. 45:243-262.

Rieseberg LH, Sinervo B, Linder CR, Ungerer MC, Arias DM. 1996b. Role of gene interactions in hybrid speciation: evidence from ancient and experimental hybrids. Science. 272:741-744.

Ruhfel BR, Gitzendanner MA, Soltis PS, Soltis DE, Burleigh JG. 2014. From algae to angiosperms – inferring the phylogeny of green plants (*Viridiplantae*) from 360 plastid genomes. BMC Evol Biol. 14:23.

Saarela JM, Graham SW. 2010. Inference of phylogenetic relationships among the subfamilies of grasses Poaceae: Poales. using meso-scale sampling of the plastid genome. Botany. 88:65-84.

Sang T, Zhong Y. 2000. Testing hybridization hypotheses based on incongruent gene trees. Syst Biol. 49:422-434.

Schumann CM, Hancock JF. 1989. Paternal inheritance of plastids in *Medicago sativa.* Theor Appl Genet. 78:863-866.

Shore JS, Mcqueen KL, Little SH. 1994. Inheritance of plastid DNA in the *Turnera ulmifolia* complex (Turneraceae). Am J Bot. 81:1636-1639.

Shore JS, Triassi M. 1998. Paternally biased cpDNA inheritance in *Turnera ulmifolia* (Turneraceae). Am J Bot. 85:328-332.

Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, Jaiswal P, Mockaitis K, Liston A, Mane SP, et al. 2010. The genome of woodland strawberry (*Fragaria vesca*). Nat Genet. 43:109-116.

Simon S, Narechania A, DeSalle R, Hadrys H. 2012. Insect phylogenomics: Exploring the source of incongruence using new transcriptomic data. Genome Biol Evol. 4:1295-1309.

Smith SA, Wilson NG, Goetz FE, Feehery C, Andrade SCS, Rouse GW, Giribet G, Dunn CW. 2011. Resolving the evolutionary relationships of molluscs with phylogenomic tools. Nature. 480:364-367.

Soltis DE, Kuzoff RK. 1995. Discordance between nuclear and chloroplast phylogenies in the *Heuchera* group (Saxifragaceae). Evolution. 49:727-742.

Soltis PS, Soltis DE, Chase MW. 1999. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. Nature. 402:402-404.

Soltis DE, Soltis PS, Chase MW, Mort ME, Albach DC, Zanis M, Savolaninen V, Hahn WH, Hoot SB, Fay MF, et al. 2000. Angiosperm phylogeny inferred from 18S rDNA, *rbcL*, and *atpB* sequences. Bot J Linn Soc. 133:381-461.

Soltis DE, Soltis PS, Endress PK, Chase MW. 2005. Phylogeny and evolution of angiosperms. Sunderland (MA): Sinauer Associates.

Soltis DE, Gitzendanner MA, Soltis PS. 2007. A 567-taxon data set for angiosperms: The challenges posed by Bayesian analyses of large data sets. Int J Plant Sci. 168:137-157.

Soltis DE, Moore MJ, Burleigh JG, Bell CD, Soltis PS. 2009. Molecular markers and concepts of plant evolutionary relationships: progress, promise, and future prospects. CRC Crit Rev Plant Sci. 28:1-15.

Soltis DE, Soltis PS. 2009. The role of hybridization in plant speciation. Annu Rev Plant Biol. 60:561-588.

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. Am J Bot. 98:704-730.

Song S, Liu L, Edwards SV, Wu S. 2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proc Natl Acad Sci USA. 109:14942-14947.

Stamatakis A, Hoover P, Rougemont J. 2008. A Rapid bootstrap algorithm for the RAxML web servers. Syst Biol. 57:758-771.

Testolin R, Cipriani G. 1997. Paternal inheritance of chloroplast DNA and maternal inheritance of mitochondrial DNA in the genus *Actinidia.* Theor Appl Genet. 94:897-903.

Tsitrone A, Kirkpatrick M, Levin DA. 2003. A model for chloroplast capture. Evolution. 57:1776-82.

Wang HC, Moore MJ, Soltis PS, Bell CD, Brockington SF, Alexandre R, Davis CC, Latvis M, Manchester SR, Soltis DE. 2009. rosids radiation and the rapid rise of angiosperm-dominated forests. Proc Natl Acad Sci USA. 106:3853-3858.

Wendel JF, Schnabel A, Seelanan T. 1995. An unusual ribosomal DNA sequence from *Gossypium gossypioides* reveals ancient, cryptic, intergenomic introgression. Mol Phylogenet Evol. 4:298-313.

Whitfield JB, Lockhart PJ. 2007. Deciphering ancient rapid radiations. Trends Ecol Evol. 22:258-265.

Xiang, QYJ, Manchester, SR, Thomas, DT, Zhang, W, Fan, C. 2005. Phylogeny, biogeography, and molecular dating of cornelian cherries (*Cornus*, Cornaceae): tracking Tertiary plant migration. Evolution. 59:1685-1700.

Xu J. 2005. The inheritance of organelle genes and genomes: patterns and mechanisms. Genome. 48:951-958.

Yang TW, Yang YA, Xiong Z. 2000. Paternal inheritance of chloroplast DNA in interspecific hybrids in the genus *Larrea* (Zygophyllaceae). Am J Bot. 87:1452-1458.

Yoder JB, Briskine R, Mudge J, Farmer A, Paape T, Steele K, Weiblen GD, Bharti AK, Zhou P, May GD, et al. 2013. Phylogenetic signal variation in the genomes of *Medicago* (Fabaceae). Syst Biol. 62:424-438.

Yu Y, Than C, Degnan JH, Nakhleh L. 2011. Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. Syst Biol. 60:138-149.

Zhang N, Zeng LP, Shan HY, Ma H. 2012. Highly conserved low-copy nuclear genes as effective markers for phylogenetic analyses in angiosperms. New Phytol. 195:923-937.

Zhong, B, Deusch, O, Goremykin, VV, Penny, D, Biggs, PJ, Atherton, RA, Nikiforova, SV, Lockhart, PJ. 2011. Systematic error in seed plant phylogenomics. Genome Biol Evol. 3:1340-1348.

Zhu XY, Chase MW, Qiu YL, Kong HZ, Dilcher DL, Li JH, Chen ZD. 2007. Mitochondrial *matR* sequences help to resolve deep phylogenetic relationships in rosids. BMC Evol Biol. 7:217.

Zou XH, Zhang FM, Zhang JG, Zang LL, Tang L, Wang J, Sang T, Ge S. 2008. Analysis of 142 genes resolves the rapid diversification of the rice genus. Genome Biol. 9:R49.

Supplementary material will be available on Dryad (http://datadryad.org**):**

Supplementary material, including all original sampling lists, alignments, and corresponding tree files.

S1. Compressed file of 8,445 potentially informative ortholog sets from Lee at al. (2011) for single-copy nuclear gene analysis.

S2. Sampling list of 22 plant taxa with genome sequences from OrthoMCL for multi-copy nuclear gene analysis.

S3. Collection of 3,748 gene family alignments from 22 plant taxa for multi-copy nuclear gene analysis.

S4. Files from the 82-taxon, 78-gene cpDNA analysis. This compressed file includes: 82-taxon, 78-gene cpDNA nucleotide, AA, and RY-coding alignments; associated AA tree, RY-coding tree, and 78 individual gene trees; trees and sorted matrices from method to remove highly variable sites.

S5. Files from the 79-taxon, 4-gene mtDNA analysis. This compressed file includes: 79-taxon, 4-gene mtDNA nucleotide, AA, and RY-coding alignments; associated AA tree, RY-coding tree, and 4 individual gene trees; trees and sorting matrices method to remove highly variable sites.

S6. Files from the 92-taxon, 5-gene nuDNA analysis. This compressed file includes: 92-taxon, 5-gene nuDNA nucleotide, AA, and RY-coding alignments; associated AA tree, RY-coding tree, and 5 individual gene trees; sorting matrices and trees from method to remove highly variable sites.

Figure Legends

Figure 1: Majority-rule consensus tree of maximum likelihood bootstrap analysis of 78-gene chloroplast matrix. These data place the COM clade with *Fabidae*. To highlight the position of the COM clade, only the *Fabidae*, COM, and *Malvidae* clades were labeled, and the COM clade is isolated from the circumscription of *Fabidae* as previously defined (Cantino et al. 2007). All the familial and ordinal names of the sampled taxa follow APG III (2009). Numbers above branches are bootstrap percentages (BS).

Figure 2: Majority-rule consensus tree of maximum likelihood bootstrap analysis of 4-gene mitochondrial matrix. The COM clade is placed with *Malvidae*.

Figure 3: Majority-rule consensus tree of maximum likelihood bootstrap analysis of 5-gene nuclear matrix. These data place the COM clade with *Malvidae*.

Figure 4: Hypothetical reticulation scenario for the origin of the COM clade from the ancestral *Fabidae* and *Malvidae* lineages. Large circles reflect the plant lineages; the small circles represent their nuclear DNA types (the red circle represents the *Fabidae* nuclear DNA and the blue one the *Malvidae* nuclear DNA), the ovals represent the chloroplast (the green ovals represent the *Fabidae* chloroplast, and the gray oval represents the chloroplast from the *Malvidae* ancestor), and the diamonds represent the mitochondria (the gray diamond represents the *Fabidae* mitochondrion and the orange diamond the *Malvidae* mitochondrion); dashed arrows represent multiple generations of backcrossing. During hybridization, the mitochondrion is maternally inherited from the *Malvidae* ancestor, and the chloroplast is paternally inherited from the *Fabidae* ancestor. After subsequent F_1 backcrosses to the *Malvidae*, the resulting generations contain chloroplasts from *Fabidae*, mitochondria from *Malvidae,* and a majority of the nuclear genes from *Malvidae*, with a smaller number from *Fabidae* (roughly 25%). The reticulate phylogeny at the bottom illustrates this hypothetical introgressive hybridization scenario and shows the phylogenetic incongruence among the three genomes for the COM clade.

Table 1 Summary of the placement of the COM clade in previous phylogenetic studies.

Genome type	Relationship[a]	Method of analysis /Support[b]	Marker[c]	Taxa Number	COM sampling	References
Chloroplast	Nr	Character-state weighting/–	*rbcL*	499	–	Chase et al. 1993
	COM + *Fabidae*	Parsimony/52% JK; BI/1.0 PP	*matK*	374	16	Hilu et al. 2003
	COM + *Fabidae*	Parsimony jackknifing/77% JK; BI/1.0 PP	*rbcL*, *atpB*, 18S rDNA	560	64	Soltis et al. 2000; Soltis et al.2007
	COM + *Fabidae*	ML/100% BS; MP/79% BS; BI/1.0 PP	81 cp	64	3	Jansen et al. 2007
	COM + *Fabidae*	ML/89% BS	*rbcL*, *atpB*, *matK*, 18S rDNA, 26S rDNA	567	59	Burleigh et al. 2009
	COM + *Fabidae*	ML/100% BS	10 cp, 2 nu	117	33	Wang et al. 2009
	COM + *Fabidae*	ML/53% BS	83 cp	86	5	Moore et al. 2010
	COM + *Fabidae*	ML/99% BS (244 taxa); ML/89% BS (87 taxa)	IR	244	14	Moore et al. 2011
	COM + *Fabidae*	ML/57% BS	11 cp, 2 nu, 4 mt	640	154	Soltis et al. 2011
	COM + *Fabida*	ML/81% BS, 70% BS, 82% BS, 69% BS (ntAll, ntNo3rd, RY, AA)	78 cp	360	9	Ruhfel et al. 2014
Mitochondrial	COM + *Malvidae*	ML/54% BS; MP/–	*matR*	174	21	Zhu et al. 2007
	COM + *Malvidae*	ML/99% BS	*atp1*, *matR*, *nad5*, *rps3*	380	26	Qiu et al. 2010
Nuclear	Nr	–	18S rDNA	233	/	Soltis et al. 1997
	Oxalidales-M	ML/55% BS	*Xdh*	247	19	Morton 2011
	COM + *Malvidae*	ML/>95% BS; BI/1.0 PP	SMC1, SMC2, MCM5, MLH1, MSH1	94	5	Zhang et al. 2012
	Malpighiales-M	GTP-ML/18% BS (136 taxa);	18,896 gene trees	136	15	Burleigh et al. 2010

		GTP-ML/75% BS (54 taxa)				
	COM + *Malvidae*	ML/>95% BS; MP/≤65% BS	nuclear genome	101	7	Lee at al. 2011

[a] Nr = not resolved; COM + *Fabidae* = COM clade was placed in *Fabidae*; COM + *Malvidae* = COM clade sister to *Malvidae*; Oxalidales-M = only Oxalidales sister to *Malvidae*; Malpighiales-M = only Malpighiales sister to *Malvidae*.

[b] JK = Jackknife value; BI = Bayesian inference; BS = Bootstrap value; PP = Posterior probabilities; GTP = Gene tree parsimony;

[c] 81 cp = 81 chloroplast genes (Jansen et al. 2007); 10 cp, 2 nu = 10 chloroplast genes, *rbcL*, *atpB*, *matK*, *psbBTNH* region (4 genes), *rpoC2*, *ndhF*, and *rps4*, and two nuclear genes, 18S rDNA and 26S rDNA; 83 cp = 83 chloroplast genes (Moore et al. 2010); IR means 25,000-bp plastid Inverted Repeat region; 11 cp, 2 nu, 4 mt = 11 chloroplast genes, *rbcL*, *atpB*, *matK*, *psbBTNH* region (4 genes), *rpoC2*, *ndhF*, *rps4* and *rps16*, and two nuclear genes, 18S rDNA and 26S rDNA, and four mitochondrial genes, *atp1*, *matR*, *nad5*, and *rps3*; ntAll, ntNo3rd, RY, AA = four different character-coding matrix; ntAll = all nucleotide positions analysis; ntNo3rd = the first and second codon positions analysis; RY = RY-coded analysis; AA = the amino acid analysis; 78 cp= chloroplast genes (Ruhfel et al. 2014).

Table 2 Results from the single-copy nuclear gene analysis based on the ortholog alignments from Lee et al. (2011). The columns list the number of orthologs (out of 8,445) that support a 'COM + *Fabidae*' topology ("*Fabidae*" column), a 'COM + *Malvidae*' topology ("*Malvidae*" column), or COM outside *Fabidae* + *Malvidae* ("*Fabidae* + *Malvidae*" column) with at least the specified amount of bootstrap (BS) support.
The %*Fabidae*, %*Malvidae*, and %(*Fabidae* + *Malvidae*) columns have the percentage of informative genes that support each placement of the COM clade.

% BS	***Fabidae***	**%*Fabidae***	***Malvidae***	**%*Malvidae***	***Fabidae* + *Malvidae***	**%(*Fabidae* + *Malvidae*)**
10	746	39	1178	61	0	0
20	449	39	704	61	0	0
30	283	38	471	62	0	0
40	171	35	321	65	0	0
50	115	36	208	64	0	0
60	68	34	131	66	0	0
70	49	36	89	64	0	0
80	29	38	48	62	0	0
90	15	35	28	65	0	0
100	3	25	9	75	0	0

Table 3 Results from the multi-copy gene tree analysis based on genome sequences of 22 land plant taxa. We calculated the reconciliation cost based on the minimum number of implied gene duplications, duplications + losses, and deep coalescence events, for 100 ML bootstrap gene trees made from 3,784 multi-copy gene alignments. In the table, we list the number of genes (out of 3,784) that have at least 50% bootstrap (BS) support for the three possible topologies based on the reconciliation cost, respectively. The %*Fabidae*, %*Malvidae* and %(*Fabidae* + *Malvidae*) columns have the percentage of informative genes that support each placement of the COM clade.

Gene Duplications

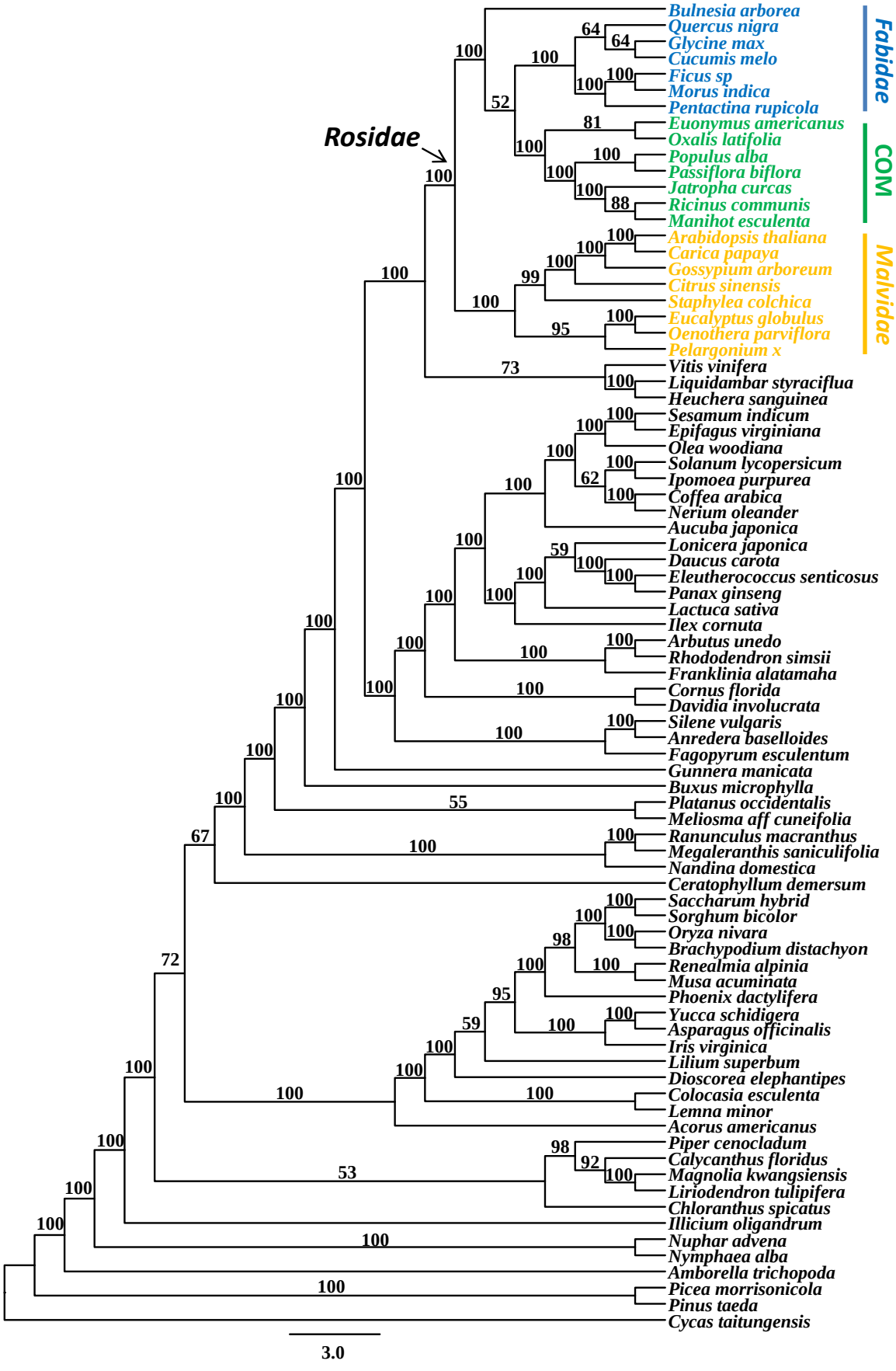
% BS	***Fabidae***	**%*Fabidae***	***Malvidae***	**%*Malvidae***	***Fabidae* + *Malvidae***	**%(*Fabidae* + *Malvidae*)**
50	38	4	973	91	62	6
60	17	2	723	94	29	4
70	12	2	515	96	11	2
80	4	1	308	98	3	1
90	2	1	155	98	1	1
100	1	6	15	94	0	0

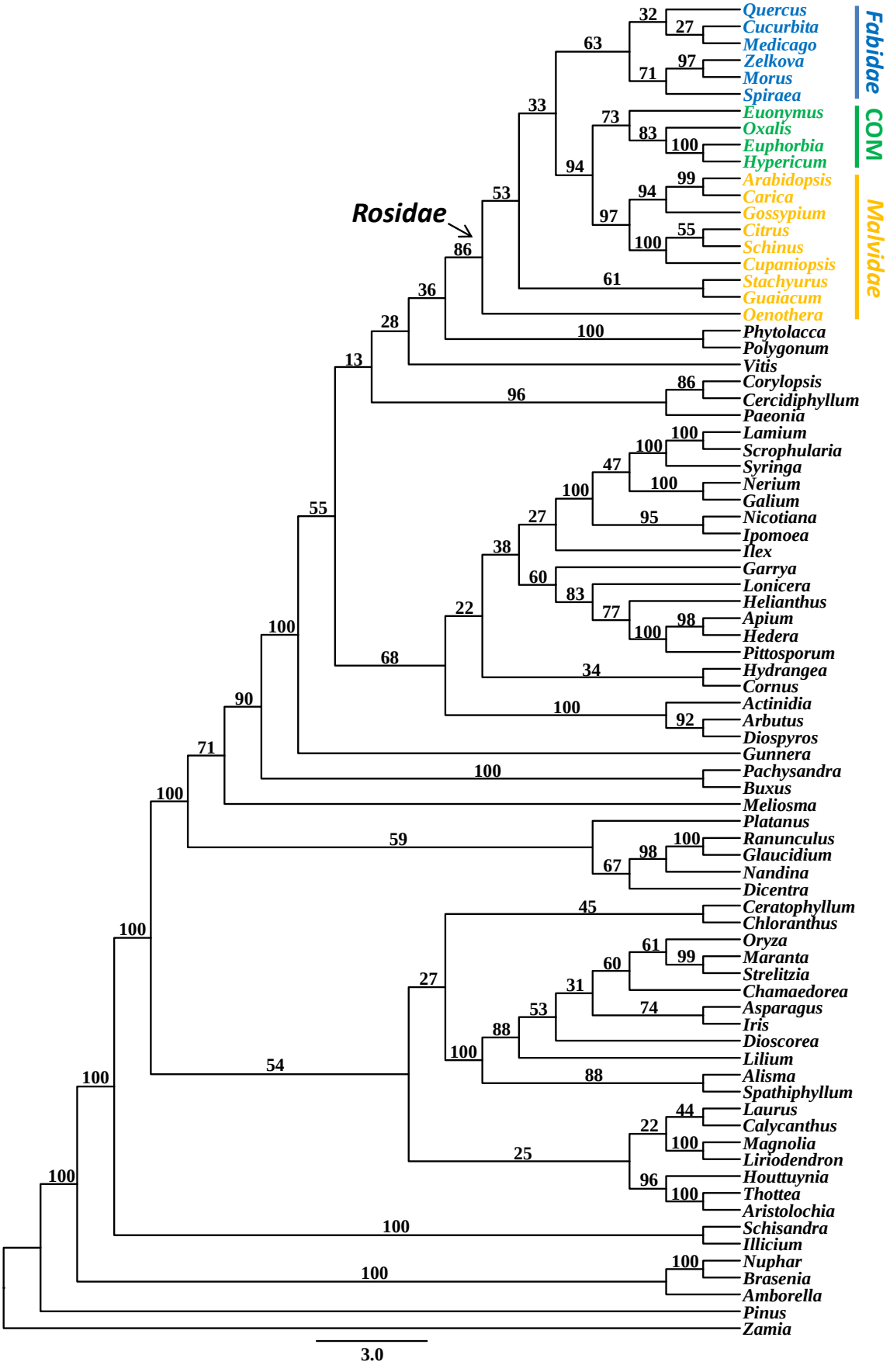
Gene Duplications + Losses

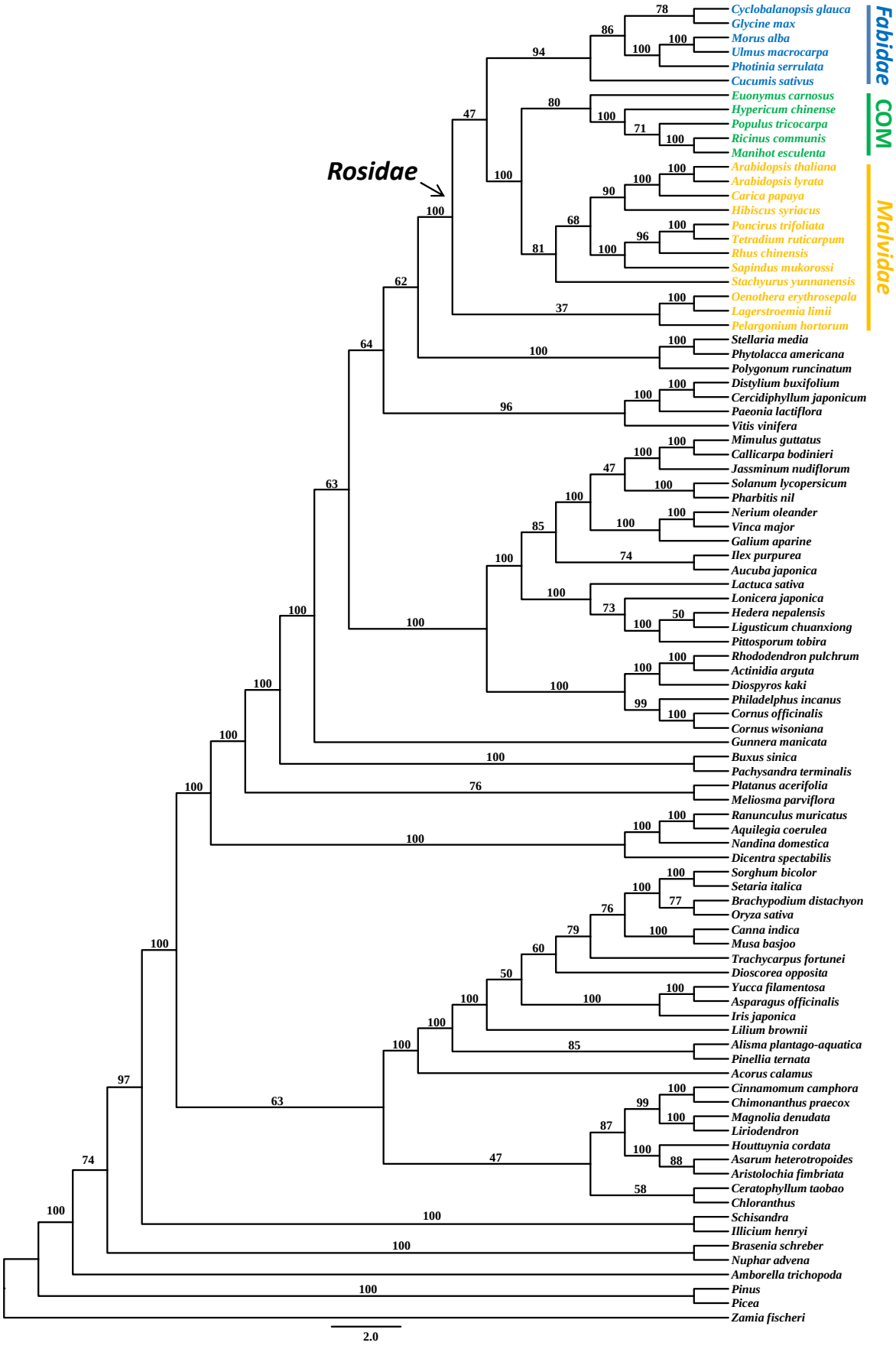
% BS	*Fabidae*	%*Fabidae*	*Malvidae*	%*Malvidae*	*Fabidae* + *Malvidae*	%(*Fabidae* + *Malvidae*)
50	446	20	1718	76	82	4
60	267	16	1390	81	47	3
70	149	12	1049	86	26	2
80	72	9	715	90	11	1
90	30	8	371	91	5	1
100	7	11	54	89	0	0

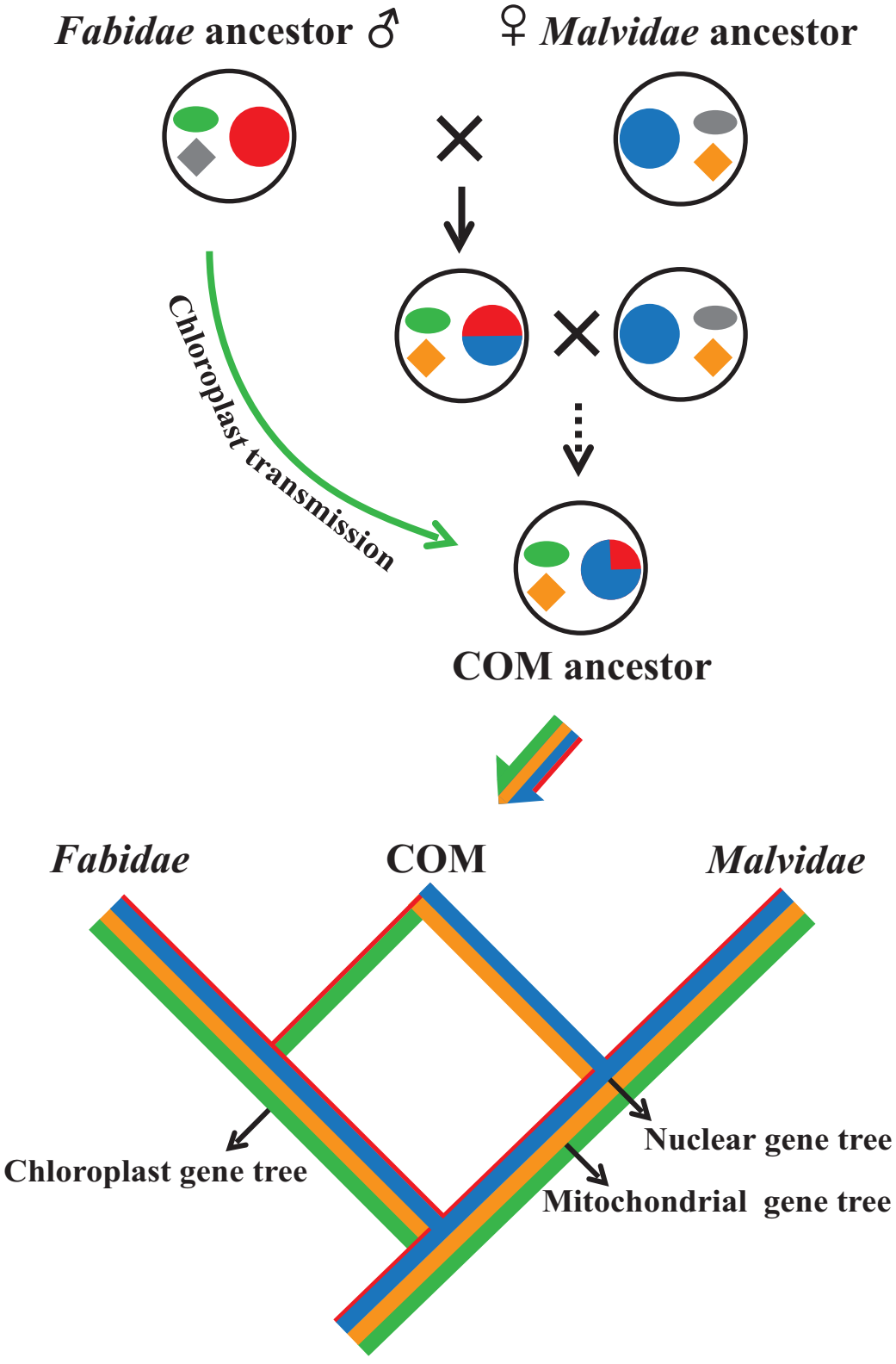
Deep Coalescence

% BS	*Fabidae*	%*Fabidae*	*Malvidae*	%*Malvidae*	*Fabidae* + *Malvidae*	%(*Fabidae* + *Malvidae*)
50	547	27	1468	71	40	2
60	358	23	1160	75	25	2
70	229	20	884	79	13	1
80	114	16	598	83	8	1
90	58	16	299	83	4	1
100	7	13	44	85	1	2


Rosidae
Fabidae
COM
Malvidae
Bulnesia arborea
Quercus nigra
Glycine max
Cucumis melo
Ficus sp
Morus indica
Pentactina rupicola
Euonymus americanus
Oxalis latifolia
Populus alba
Passiflora biflora
Jatropha curcas
Ricinus communis
Manihot esculenta
Arabidopsis thaliana
Carica papaya
Gossypium arboreum
Citrus sinensis
Staphylea colchica
Eucalyptus globulus
Oenothera parviflora
Pelargonium x
Vitis vinifera
Liquidambar styraciflua
Heuchera sanguinea
Sesamum indicum
Epifagus virginiana
Olea woodiana
Solanum lycopersicum
Ipomoea purpurea
Coffea arabica
Nerium oleander
Aucuba japonica
Lonicera japonica
Daucus carota
Eleutherococcus senticosus
Panax ginseng
Lactuca sativa
Ilex cornuta
Arbutus unedo
Rhododendron simsii
Franklinia alatamaha
Cornus florida
Davidia involucrata
Silene vulgaris
Anredera baselloides
Fagopyrum esculentum
Gunnera manicata
Buxus microphylla
Platanus occidentalis
Meliosma aff cuneifolia
Ranunculus macranthus
Megaleranthis saniculifolia
Nandina domestica
Ceratophyllum demersum
Saccharum hybrid
Sorghum bicolor
Oryza nivara
Brachypodium distachyon
Renealmia alpinia
Musa acuminata
Phoenix dactylifera
Yucca schidigera
Asparagus officinalis
Iris virginica
Lilium superbum
Dioscorea elephantipes
Colocasia esculenta
Lemna minor
Acorus americanus
Piper cenocladum
Calycanthus floridus
Magnolia kwangsiensis
Liriodendron tulipifera
Chloranthus spicatus
Illicium oligandrum
Nuphar advena
Nymphaea alba
Amborella trichopoda
Picea morrisonicola
Pinus taeda
Cycas taitungensis
3.0


Rosidae
Fabidae
COM
Malvidae
Quercus
Cucurbita
Medicago
Zelkova
Morus
Spiraea
Euonymus
Oxalis
Euphorbia
Hypericum
Arabidopsis
Carica
Gossypium
Citrus
Schinus
Cupaniopsis
Stachyurus
Guaiacum
Oenothera
Phytolacca
Polygonum
Vitis
Corylopsis
Cercidiphyllum
Paeonia
Lamium
Scrophularia
Syringa
Nerium
Galium
Nicotiana
Ipomoea
Ilex
Garrya
Lonicera
Helianthus
Apium
Hedera
Pittosporum
Hydrangea
Cornus
Actinidia
Arbutus
Diospyros
Gunnera
Pachysandra
Buxus
Meliosma
Platanus
Ranunculus
Glaucidium
Nandina
Dicentra
Ceratophyllum
Chloranthus
Oryza
Maranta
Strelitzia
Chamaedorea
Asparagus
Iris
Dioscorea
Lilium
Alisma
Spathiphyllum
Laurus
Calycanthus
Magnolia
Liriodendron
Houttuynia
Thottea
Aristolochia
Schisandra
Illicium
Nuphar
Brasenia
Amborella
Pinus
Zamia
3.0


Rosidae
Fabidae
COM
Malvidae
Cyclobalanopsis glauca
Glycine max
Morus alba
Ulmus macrocarpa
Photinia serrulata
Cucumis sativus
Euonymus carnosus
Hypericum chinense
Populus tricocarpa
Ricinus communis
Manihot esculenta
Arabidopsis thaliana
Arabidopsis lyrata
Carica papaya
Hibiscus syriacus
Poncirus trifoliata
Tetradium ruticarpum
Rhus chinensis
Sapindus mukorossi
Stachyurus yunnanensis
Oenothera erythrosepala
Lagerstroemia limii
Pelargonium hortorum
Stellaria media
Phytolacca americana
Polygonum runcinatum
Distylium buxifolium
Cercidiphyllum japonicum
Paeonia lactiflora
Vitis vinifera
Mimulus guttatus
Callicarpa bodinieri
Jassminum nudiflorum
Solanum lycopersicum
Pharbitis nil
Nerium oleander
Vinca major
Galium aparine
Ilex purpurea
Aucuba japonica
Lactuca sativa
Lonicera japonica
Hedera nepalensis
Ligusticum chuanxiong
Pittosporum tobira
Rhododendron pulchrum
Actinidia arguta
Diospyros kaki
Philadelphus incanus
Cornus officinalis
Cornus wisoniana
Gunnera manicata
Buxus sinica
Pachysandra terminalis
Platanus acerifolia
Meliosma parviflora
Ranunculus muricatus
Aquilegia coerulea
Nandina domestica
Dicentra spectabilis
Sorghum bicolor
Setaria italica
Brachypodium distachyon
Oryza sativa
Canna indica
Musa basjoo
Trachycarpus fortunei
Dioscorea opposita
Yucca filamentosa
Asparagus officinalis
Iris japonica
Lilium brownii
Alisma plantago-aquatica
Pinellia ternata
Acorus calamus
Cinnamomum camphora
Chimonanthus praecox
Magnolia denudata
Liriodendron
Houttuynia cordata
Asarum heterotropoides
Aristolochia fimbriata
Ceratophyllum taobao
Chloranthus
Schisandra
Illicium henryi
Brasenia schreber
Nuphar advena
Amborella trichopoda
Pinus
Picea
Zamia fischeri
2.0


Fabidae ancestor ♂
♀ *Malvidae* ancestor
Chloroplast transmission
COM ancestor
Fabidae
COM
Malvidae
Chloroplast gene tree
Nuclear gene tree
Mitochondrial gene tree