Molecular Phylogenetics and Evolution 105 (2016) 166–176

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Phylogenomic analyses of large-scale nuclear genes provide new insights into the evolutionary relationships within the rosids

Lei Zhao [a,b], Xia Li [a], Ning Zhang [c], Shu-Dong Zhang [a], Ting-Shuang Yi [a], Hong Ma [d], Zhen-Hua Guo [a,*], De-Zhu Li [a,*]

[a] *Plant Germplasm and Genomics Center, Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, Yunnan 650201, China*
[b] *Kunming College of Life Sciences, University of Chinese Academy of Sciences, Kunming, Yunnan 650201, China*
[c] *Department of Botany, National Museum of Natural History, MRC 166, Smithsonian Institution, Washington, DC 20013-7012, USA*
[d] *Ministry of Education Key Laboratory of Biodiversity Sciences and Ecological Engineering, Institute of Plant Biology, Institute of Biodiversity Sciences, Center for Evolutionary Biology, School of Life Sciences, Fudan University, Shanghai 200032, China*

ARTICLE INFO

Article history:
Received 28 November 2015
Revised 6 June 2016
Accepted 27 June 2016
Available online 28 June 2016

Keywords:
Coalescence
Rosids phylogenomics
Nuclear genes
Gene tree
Species tree

ABSTRACT

The Rosids is one of the largest groups of flowering plants, with 140 families and ~70,000 species. Previous phylogenetic studies of the rosids have primarily utilized organelle genes that likely differ in evolutionary histories from nuclear genes. To better understand the evolutionary history of rosids, it is necessary to investigate their phylogenetic relationships using nuclear genes. Here, we employed large-scale phylogenomic datasets composed of nuclear genes, including 891 clusters of putative orthologous genes. Combined with comprehensive taxon sampling covering 63 species representing 14 out of the 17 orders, we reconstructed the rosids phylogeny with coalescence and concatenation methods, yielding similar tree topologies from all datasets. However, these topologies did not agree on the placement of Zygophyllales. Through comprehensive analyses, we found that missing data and gene tree heterogeneity were potential factors that may mislead concatenation methods, in particular, large amounts of missing data under high gene tree heterogeneity. Our results provided new insights into the deep phylogenetic relationships of the rosids, and demonstrated that coalescence methods may effectively resolve the phylogenetic relationships of the rosids with missing data under high gene tree heterogeneity.

1. Introduction

The rosids is one of the most diverse lineages of flowering plants, containing 17 orders, which in turn comprise 140 families and ca. 70,000 species, exhibiting remarkable morphological and ecological diversities (APG II, 2003; APG III, 2009; Chase et al., 1993; Magallon et al., 1999). The rosids consists of an unusually heterogeneous group with respect to habitat and life form, with member species occurring as herbs, trees, aquatics and succulents. Some members are significant cash crops (e.g., Fabaceae, Rosaceae and Brassicaceae), and others are important forest trees (e.g., Betulaceae, Fagaceae and Sapindaceae). In previous studies based on chloroplast and mitochondrial genes, the rosids has been divided into two major clades (Fig. 1a): (i) the fabids, which contains the nitrogen-fixing clade (Cucurbitales, Fagales, Fabales and Rosales), the COM clade (Celastrales, Oxalidales and Malpighiales) and Zygophyllales; and (ii) the malvids, which includes Brassicales, Malvales, Sapindales, Crossosomatales, Picramniales, Huerteales, Geraniales and Myrtales (APG IV, 2016; Hilu et al., 2003; Judd and Olmstead, 2004; Moore et al., 2011; Qiu et al., 2010; Soltis et al., 2011; Wang et al., 2009; Zhu et al., 2007).

Despite these prior works, the positions of some clades, the COM clade, Geraniales, Myrtales and Zygophyllales remained uncertain (APG IV, 2016; Maia et al., 2014; Morton, 2011; Sun et al., 2015). The COM clade, as circumscribed by these two studies (Endress and Matthews, 2006; Zhu et al., 2007), was sister to the nitrogen-fixing clade of the fabids according to some previous studies based on chloroplast genes (Burleigh et al., 2009; Hilu et al., 2003; Jansen et al., 2007; Moore et al., 2010; Ruhfel et al., 2014; Soltis et al., 2007, 2011, 2000; Wang et al., 2009). Subsequently, relying on mitochondrial (Qiu et al., 2010; Zhu et al., 2007) and nuclear genes (Burleigh et al., 2011; Duarte et al., 2010; Lee et al., 2011; Maia et al., 2014; Xi et al., 2014; Zhang

---

* Corresponding authors at: Kunming Institute of Botany, Chinese Academy of Sciences, 132 Lanhei Road, Kunming, Yunnan 650201, China.

E-mail addresses: guozhenhua@mail.kib.ac.cn (Z.-H. Guo), dzl@mail.kib.ac.cn (D.-Z. Li).

http://dx.doi.org/10.1016/j.ympev.2016.06.007

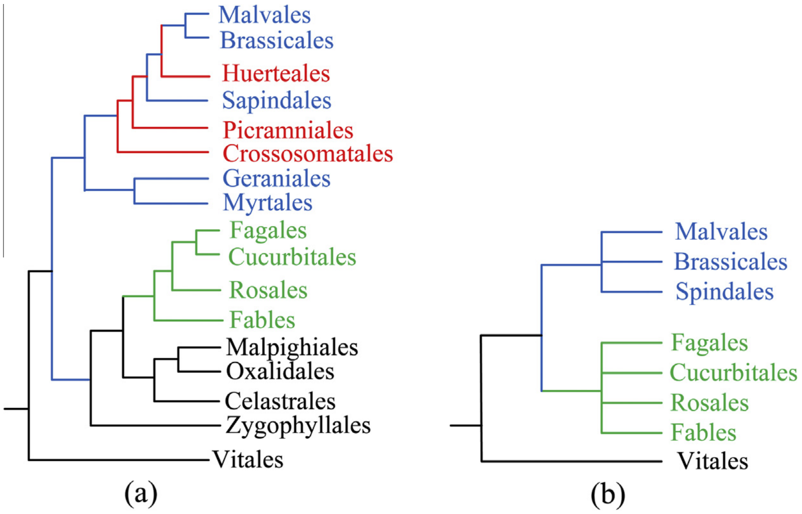

Fig. 1. The phylogeny of 17 orders within the rosids is redrawn from Wang et al. (2009), and three orders of no species sampling are in red (a). A cladogram depicts the phylogenetic relationships of eight orders, and this tree is taken as one of three rules to exclude the "problematic" gene tree (b). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

et al., 2012), other studies placed the COM clade as a part of the malvids. In addition, while the COM clade formed a monophyletic group, within it M-O (Qiu et al., 2010; Ruhfel et al., 2014; Soltis et al., 2011, 2000; Wang et al., 2009; Wu et al., 2014; Zhu et al., 2007), M-C (Burleigh et al., 2009; Moore et al., 2011; Zhang and Simmons, 2006) and O-C (Hilu et al., 2003; Moore et al., 2010; Ruhfel et al., 2014; Sun et al., 2016) were respectively supported as sister groups by different studies (Sun et al., 2015). With respect to the placements of Geraniales and Myrtales, in some studies using mitochondrial genes, they were supported as successive sisters to the remaining rosids (APG II, 2003; Hilu et al., 2003; Qiu et al., 2010; Zhu et al., 2007). However, in other studies based on chloroplast genes, they were placed in the malvids (APG IV, 2016; Jansen et al., 2007; Ruhfel et al., 2014; Soltis et al., 2011; Wang et al., 2009). For the position of Zygophyllales, the group has been placed in the malvids (Maia et al., 2014; Qiu et al., 2010; Ruhfel et al., 2014) or the fabids (Hilu et al., 2003; Ruhfel et al., 2014; Soltis et al., 2011; Wang et al., 2009). Recently, a few studies have investigated the positions of some uncertain orders by using nuclear genes, albeit with limited taxon sampling. They provided supports for grouping of some COM orders with malvids (Finet et al., 2010; Lee et al., 2011; Sun et al., 2015; Xi et al., 2014; Zhang et al., 2012) and placing Myrtales (or together with Geraniales) as sister to the remaining rosids (Myburg et al., 2014; Sun et al., 2015; Wang et al., 2014; Zeng et al., 2014).

Currently, due to increased affordability, high-throughput sequencing technologies have been widely employed for genome and transcriptome sequencing (Reuter et al., 2015). They allow data on hundreds or thousands of single or low copy nuclear genes to be collected for inferring species relationships (Lemmon and Lemmon, 2013; Wen et al., 2015; Zimmer and Wen, 2015). However, with such large, genome-scale datasets, phylogenetic conflicts among genes become evident, resulting in phenomena such as gene tree-species tree discordance (Degnan and Rosenberg, 2009; Salichos and Rokas, 2013; Szollosi et al., 2015). Therefore, simply increasing the number of gene sequences does not always resolve phylogenetic incongruences (Kimball and Braun, 2014; Nater et al., 2015; Philippe et al., 2011). Additionally, revisitations of previously published phylogenomic datasets using different analytical methods often produce conflicting results (Simmons and Gatesy, 2015; Springer and Gatesy, 2014, 2016; Tarver et al., 2016), indicating that the choice of analytical methods is an important consideration for the phylogenetic studies (Roch and Warnow, 2015).

In traditional phylogenetic analyses, multiple genes are concatenated as a supermatrix for inferring evolutionary relationships (de Queiroz and Gatesy, 2007). This concatenation method has been widely employed in numerous phylogenomic studies of animals (Jarvis et al., 2014; Kocot et al., 2011), plants (Wickett et al., 2014; Zeng et al., 2014) and fungi (Ebersberger et al., 2012; Spatafora and Bushley, 2015). Concatenation methods assume that all genes have the same evolutionary history, ignoring or downplaying inevitable evolutionary processes, such as incomplete lineage sorting (ILS), horizontal gene transfer (HGT), and gene duplication and loss (GDL) (Degnan and Rosenberg, 2009; Knowles, 2009; Nakhleh, 2013). Species tree estimation from large multi-locus datasets could be complicated by these biological processes (Edwards, 2009; Kutschera et al., 2014; Lambert et al., 2015; Som, 2015), because they cause gene tree heterogeneity, which are not explicitly accounted for by concatenation methods (Knowles, 2009; Nosenko et al., 2013; Salichos and Rokas, 2013; Szollosi et al., 2015). Other issues that may complicate phylogenetic estimation are substitution saturation, long-branch attraction (LBA) and missing data, although these issues are not necessarily exclusive to concatenation methods (Liu et al., 2015b; Roure et al., 2013; Whelan et al., 2015; Xi et al., 2016). As a result of these issues, concatenation methods may introduce significant errors or produce highly supported but incorrect species tree topologies (Giarla and Esselstyn, 2015; Linkem et al., 2016; Roch and Steel, 2014; Zhong et al., 2013).

Recently, many coalescence methods have been developed to address these problems (Knowles, 2009; Liu et al., 2015a; Szollosi et al., 2015). The first type of methods is termed co-estimation methods, e.g., BEST (Liu, 2008) and *BEAST (Heled and Drummond, 2010), which simultaneously infer gene trees and the underlying species tree. These methods have outstanding accuracy, but are computationally demanding for large datasets (Leache and Rannala, 2011; Mirarab et al., 2014b). The second type of methods is called single-site methods and they use single nucleotide polymorphisms (SNPs) to infer species trees. Examples of software that implement this type of methods are SNAPP (Bryant et al., 2012) and SVDquartets (Chou et al., 2015). The third type of methods is called summary methods. Relying on the multi-species coalescent model, methods in this class produce statistically