密级:（涉密论文填写密级，公开论文不填写）

中国科学院大学
University of Chinese Academy of Sciences

博士学位论文

蔷薇类 COM 支深层次系统关系冲突探讨：系统发育基因组学方法

作者姓名：孙 苗

指导教师：陈之端 研究员

中国科学院植物研究所

学位类别：博士

学科专业：植物学

培养单位：中国科学院植物研究所

2014 年 5 月

Exploring deep phylogenetic incongruence of the COM clade in *Rosidae*: Phylogenomics approach

By

Miao Sun

A Dissertation Submitted to

University of Chinese Academy of Sciences

In partial fulfillment of the requirement

For the degree of

Doctor of Natural Science

Institute of Botany, the Chinese Academy of Sciences

May, 2014

关于学位论文使用授权的说明

本人完全了解中国科学院植物研究所有关保留、使用学位论文的规定，即：植物研究所有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以提供目录检索以及公布论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存论文。

(保密的论文在解密后遵守此规定)

学位论文作者签名：

年　　月　　日

经指导教师同意，本学位论文属于保密，在　　年解密后适用本授权书。

指导教师签名：		**学位论文作者签名：**	
解　密　时　间：	年　　月　　日		

各密级的最长保密年限及书写格式规定如下：

内部　5 年（最长 5 年，可少于 5 年）

秘密★10 年（最长 10 年，可少于 10 年）

机密★20 年（最长 20 年，可少于 20 年）

学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，进行研究工作所取得的成果。除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人创作的、已公开发表或者没有公开发表的作品的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。本学位论文原创性声明的法律责任由本人承担。

学位论文作者签名：

年　　月　　日

本论文得到国家自然科学基金委面上项目基金（NNSF 31270268）、国家重点基础研究发展计划（No. 2014CB954101）及中国科学院外国专家特聘研究员计划（No. 2011T1S24）的联合资助，特此致谢！

摘　要

与日俱增的基因组数据不仅能够解决生命之树分支上的疑难节点，而且还有助于揭示诸如基因水平转移、不完全谱系筛选、古杂交及渐渗等复杂的生物进化过程。COM 支（Celastrales-Oxalidales-Malpighiales clade）是被子植物蔷薇类（*Rosidae*）内系统发育关系尚未确定的重要分支之一。前人分别用不同基因组和形态特征的数据不同程度地支持 COM 支分别与蔷薇类内两大亚支 *Fabidae*（豆类）和 *Malvidae*（锦葵类）近缘。为了进一步确认 COM 支的系统位置，并探讨上述系统关系分歧的原因，本研究设计了代表植物三个基因组的、且取样近等同的矩阵：叶绿体 82 类群 78 基因矩阵、线粒体 79 类群 4 基因矩阵以及核 92 类群 5 基因矩阵；然后对上述三个矩阵运用了多种性状编码和数据筛除的分析方法（如，RY 编码、快速进化位点移除法）来检测 COM 支系统位置的冲突是与取样偏差或系统误差有关，还是与生物进化过程有关。之后，基于 COM 支所有可能的系统位置及核基因的双亲遗传和非连锁的特性，我们又对两个核基因组矩阵（8,445 个单拷贝矩阵和 3,748 个多拷贝矩阵）进行了分析以探讨 COM 支在核基因组内系统位置以及可能与此关联的进化事件。最后，三个基因组小矩阵的分析结果，重现了 COM 支两种冲突的系统位置，并表明此冲突是客观存在的，不是由取样偏差或系统误差造成的。单拷贝和多拷贝核基因组矩阵的分析结果与前人基于线粒体和核基因的结论一致，也支持 COM 支与 *Malvidae* 更近缘。而且，在两个核基因组矩阵分析中，我们均检测到了一定比例的少数基因支持 COM 支与 *Fabidae* 近缘的系统发育信号，但对 *Fabidae* 与 *Malvidae* 聚类之后，COM 支再为其姊妹群的系统位置几乎没有支持。最后，从核基因组内对 COM 支系统位置支持基因的比例来看，基本排除了不完全谱系筛选的造成冲突的可能性，暗示了三个类群间可能存在着 COM 支以 *Fabidae* 和 *Malvidae* 为亲本的遗传关系。

综上所述，本研究认为 COM 支与 *Malvidae* 的姊妹关系代表了 COM 支真实系统发育关系，不支持取样偏差或者系统误差造成 COM 支在三个植物基因组间系统位置冲突，并进一步揭示了蔷薇类早期快速辐射分化过程中，*Fabidae* 和 *Malvidae* 谱系的祖先间可能发生了古杂交并随叶绿体基因组渐渗的进化事件，

从而造成了关于 COM 支的叶绿体与核、线粒体基因树间的冲突，最终形成了核基因组内两种相互冲突且代表亲本遗传信息的系统发育信号共存的格局。尽管有待于扩大类群和核基因组的取样来进一步确认这些网状进化事件的发生，但本研究不仅为验证、解决不同基因组间生命之树其他分支系统位置冲突的问题提供了借鉴和示范，而且还着重体现了基因组数据在揭示生命之树深层次系统发育关系的冲突及复杂生物过程方面的重要意义。

关键词：古杂交；渐渗；不完全谱系筛选；COM 支；冲突；系统发育基因组学

Exploring deep phylogenetic incongruence of the COM clade in *Rosidae*: Phylogenomics approach

Miao Sun (Botany)
Directed by Prof. Zhiduan Chen

Abstract

Analysis of large data sets can help resolve difficult nodes in the tree of life and also concomitantly 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* (suggested by chloroplast genes) or *Malvidae* (suggested by mitochondrial and nuclear genes, as well as morphological data). To elucidate the underlying cause of this phylogenetic discordance, we assembled taxonomically comparable multi-gene matrices of chloroplast, mitochondrial, and nuclear sequences (82 taxa for 78 chloroplast genes, 79 taxa for 4 mitochondrial genes, and 92 taxa for 5 nuclear genes), as well as large single- and multi-copy nuclear gene data sets (8,445 single-copy ortholog sets and 3,748 multi-copy nuclear gene families). 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, not due to systematic biases or sampling errors. 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, remaining ancient introgressive hybridization as a plausible explanation for the conflict among genes.

In summary, our analyses demonstrate that the placement of COM clade with *Malvidae* better reflects organismal phylogeny, the conflicting phylogenetic

placements for COM clade might be caused by ancient hybridization and chloroplast transmission occurred during the early and rapid radiation of *Rosidae*, and consequently resulted in conflict between chloroplast and mitochondrial gene trees, as well as a mixture of two underlying signals derived from its ancestral parents in the nuclear genome. Although greater taxon and nuclear genome sampling are necessary to evaluate such hypothesis fully, our study provides an example for examination of other deep nodes of the tree of life where conflict occurs among data sets from different subcellular compartments, and also emphasizes the importance of genomic data sets for revealing deep incongruence and potentially complex patterns of evolution in organismal phylogeny.

Key Words: Hybridization; introgression; incomplete lineage sorting; COM clade; incongruence; phylogenomics

目 录

第一章 前言

物种及其以上分类阶元（属、科、目、纲、门、界）生物类群或谱系之间的系统发育关系是理解和认识地球生物起源和进化的基础，是系统与进化生物学和其他相关学科需要解决的核心问题之一（Rokas et al. 2003; Delsuc et al. 2005; Whitfield and Lockhart 2007; Zou and Ge 2008; Zhang et al. 2012）。早在 19 世纪中叶，Darwin（1859）就认为地球上的一切生命形式都有一个共同的起源，物种之间有着一种树状结构的关联，即“the great Tree of Life”。此后，诸多生物学家、系统学家都致力于利用形态、分子、化学等方面的证据来追溯物种间的亲缘关系和进化格局（见 Hong et al. 2008），其中分子系统学证据在生物学领域逐渐受到青睐并被广泛应用。尤其是近些年来，随着 PCR 和测序技术的日趋发展进步，植物系统发育关系的研究经历了由最初的若干类群单基因测序到更大规模类群的多基因、转录组，再到整个基因组规模的系统发育分析（如，Chase et al. 1993; Goloboff et al. 2009; Jansen et al. 2007; Moore et al. 2010, 2011; Burleigh et al. 2011; Lee et al. 2011; Smith et al. 2011; Ruhfel et al. 2014）。可见，人们在了解和认识植物的起源和分化、澄清生命之树各大分支的亲缘关系方面取得了重大的进展（APG 1998; APG II 2003; Judd and Olmstead 2004; Soltis and Soltis 2004; Chase et al. 2006; Frohlich and Chase 2007; APG III 2009; Soltis et al. 2005, 2007, 2008, 2009, 2010, 2011）。如，单子叶植物分支（*Monocotyledonae*, monocots；Chase et al. 2000; Jerrold et al. 2004; Graham et al. 2006; Givnish et al. 2006, 2010; Saarela et al. 2008），菊类分支（*Asteridae*, asterids；Baldwin 1992; Olmstead et al. 2000; Albach et al. 2001; Bremer et al. 2001, 2004; Hilu et al. 2003），蔷薇类分支（*Rosidae*, rosids; Hilu et al. 2003; Jasen et al. 2007; Soltis et al. 2007, 2011; Zhu et al. 2007; Wang et al. 2009; Moore et al. 2010; Qiu et al. 2010）。而另一方面，人们在重建生命之树过程中也会遇到一些棘手的问题。比如，随机的一棵基因树可能与普遍接受的物种树不一致（Degnan and Rosenberg 2006; Huang and Knowles 2009）；而联合多个基因的系统发育分析又会因各个基因具有不同的进化历史而可能产生高度支持但错误的拓扑关系（Mossel and Vigoda 2005; Kubatko and Degnan 2007; Matsen and Steel 2007; Beiko et al. 2008; Penney et al. 2008; Salichos and Rokas 2013）。如

今，与日俱增的基因组数据及针对于此的生物信息分析方法不仅是解决生命之树重建过程中上述疑难问题的有效工具（Dunn et al. 2008; Lee et al. 2011; Smith et al. 2011; Simon et al. 2012; Yoder et al. 2013），也是揭示潜藏在系统发育关系冲突背后生物过程（如，基因重复和丢失 gene duplication and loss、不完全谱系筛选 incomplete lineage sorting、基因水平转移 horizontal gene transfer、基因重组 gene recombination、杂交 hybridization）的强而有力的手段（Goodman et al. 1979; Hudson 1983; Doyle 1992; Maddison 1997; Degnan and Rosenberg 2009; Cui et al. 2013; Oliver 2013）。

尽管目前被子植物系统发育关系的基本骨架已经确立，但在此大尺度上的研究工作主要是依据叶绿体基因组数据来开展的，而在此尺度上线粒体基因的研究较少（Zhu et al. 2007; Qiu et al. 2010），核基因由于其自身的重复/丢失及多拷贝的限制也未能被广泛应用（Soltis et al. 1997; Morton 2011; Zhang et al. 2012; Zeng et al. 2014），多与叶绿体片段结合使用（如，Wang et al. 2009; Soltis et al. 2011）。下一代测序（next-generation sequencing, NGS）技术的诞生和发展，使得利用核基因数据评估、验证叶绿体基因组及其他数据所建立的系统发育框架和其间冲突的关系成为可能（如，Duarte et al. 2010; Shulaev et al. 2010; Burleigh et al. 2011; Lee et al. 2011; Zhang et al. 2012）。

另一方面，在植物进化过程中渐渗杂交（introgressive hybridization）事件很普遍（Rieseberg and Soltis 1991），且起着非常关键的作用（Tsitrone et al. 2003; Linder and Rieseberg 2004; Chang et al. 2011）。早在 20 多年前，在被子植物中叶绿体基因渐渗现象的报道就有 100 多例（Rieseberg et al. 1996a）。越近缘或分化时间越短的物种间（快速辐射分化的类群间），发生杂交或基因渐渗事件就会越频繁；杂交在其本质上是基因渐渗的延伸（Simpson 2012）。双亲遗传的核基因树与细胞质单亲遗传的叶绿体（或线粒体）基因树之间的冲突往往是物种杂交造成的。在短时间内的快速辐射进化过程中，不完全谱系筛选（也称 deep coalescence）同样也极易发生，造成基因树与物种树冲突（Maddison 1997; Page and Charleston 1998; Maddison and Knowles 2006）。特别是分化时间间隔越短，有效种群越大，不完全谱系筛选事件发生的可能性越大（Pamilo and Nei 1998; Rokas and Carroll 2006; Galtier and Daubin 2008; Degnan and Rosenberg 2009）。如此一来，在植物系统发育关系重建中则不乏叶绿体与核基因树相冲突的研究报

道（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）。

在通常情况下，核基因与叶绿体基因重建的系统发育关系是相吻合的。但被子植物蔷薇类中的 COM 支（Celastrales-Oxalidales-Malpighiales clade; Endress and Matthews 2006）却是例外，因为无论是形态，还是线粒体与核基因的数据获得的 COM 支的系统位置都与叶绿体基因的分析结果相冲突（表 1-1；详见本章 1.1.2）。由于前人研究的取样策略与分析方法大不相同（表 1-1），故我们并不清楚 COM 支系统位置的冲突是由实验或分析过程中的系统误差所为，还是生物过程导致；而且，核基因组内 COM 支的系统位置如何也尚不清楚。因此，对被子植物蔷薇类的这一重要分支—COM 支的系统发育关系展开深度的剖析，分析其冲突的原因，并解决此冲突，是本研究所关注的核心问题。

以下我们将阐述 COM 支的名称来源，简要回顾 COM 支系统发育关系的研究历史，总结可能造成系统关系冲突的原因及对应的分析策略和解决办法，并在此基础之上，提出 COM 支系统位置冲突的解决方案和本研究目的、意义。

1.1 COM 支的系统发育研究概述

1.1.1 COM 支的概念

COM 支的概念是由 Endress 和 Matthews（2006）在为蔷薇类分子系统发育关系提供形态性状支持而对该类群花结构特征进行调查研究时首次提出的。随后，此概念在后来的研究中一直沿用（如，Zhu et al. 2007; Zhang et al. 2012; Ruhfel et al. 2014）。COM 支隶属于蔷薇类，包括卫矛目（Celastrales）、酢浆草目（Oxalidales）和金虎尾目（Malpighiales），约有 48-50 科，870 属，19,105 种，占蔷薇类的三分之一（图 1-1; APG III 2009; Steven 2001 onwards）。在分子系统发育研究中，COM 支的单系性得到了叶绿体、线粒体和核基因单独或联合分析的一致支持（Hilu et al. 2003; Soltis et al. 2005, 2011; Jansen et al. 2007; Zhu et al. 2007; Burleigh et al. 2009; Moore et al. 2011；Qiu et al. 2010; Zhang et al. 2012）；在花形态结构研究中，珠心（nucellus）较薄和珠被绒毡层（endothelium）的出现是支持 COM 支单系性的共近裔特征（Endress and Matthews 2006; Endress et al. 2013）。

表 1-1 历史研究中 COM 支系统位置概览

Table1-1 Summary of the placement of the COM clade in previous phylogenetic studies

基因组类型	系统位置 [a]	分析方法/支持率 [b]	分子标记 [c]	类群总取样	COM 支取样	参考文献
叶绿体	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	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
线粒体	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

	Nr	–	18S rDNA	233	/	Soltis et al. 1997
	Oxalidales-M	ML/55% BS	*Xdh*	247	19	Morton 2011
核	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); GTP-ML/75% BS (54 taxa)	18,896 gene trees	136	15	Burleigh et al. 2010
	COM + *Malvidae*	ML/>95% BS; MP/≤65% BS	nuclear genome	101	7	Lee at al. 2011

注：a、Nr = 未得到解决；COM + *Fabidae* = COM 支与 *Fabidae* 聚类；COM + *Malvidae* = COM 支与 *Malvidae* 聚类；Oxalidales-M = 酢浆草目与 *Malvidae* 成姊妹关系；Malpighiales-M = 金虎尾目与 *Malvidae* 成姊妹关系；

b、JK = Jackknife value; BI = Bayesian inference; BS = Bootstrap value; PP = Posterior probabilities; GTP = Gene tree parsimony，基因树简约法；

c、81 cp = 81 个叶绿体基因（Jansen et al. 2007）；10 cp, 2 nu = 10 个叶绿体基因，包括 *rbcL*、*atpB*、*matK*、*psbBTNH* 区域（4 个基因）、*rpoC2*、*ndhF*、*rps4* 以及两个核核糖体基因 18S rDNA 和 26S rDNA；83 cp = 83 个叶绿体基因（Moore et al. 2010）；IR = 25,000 bp 的质体反向重复区；11 cp, 2 nu, 4 mt = 11 个叶绿体基因，包括 *rbcL*、*atpB*、*matK*、*psbBTNH* 区域（4 个基因）、*rpoC2*、*ndhF*、*rps4*、*rps16*，以及两个核核糖体基因 18S rDNA 和 26S rDNA，四个线粒体基因，即 *atp1*、*matR*、*nad5* 和 *rps3*；ntAll, ntNo3rd, RY, AA = 分别为核苷酸，第一、二密码子，RY 编码和氨基酸四种不同性状编码矩阵；78 cp= 78 个叶绿体基因（Ruhfel et al. 2014）。

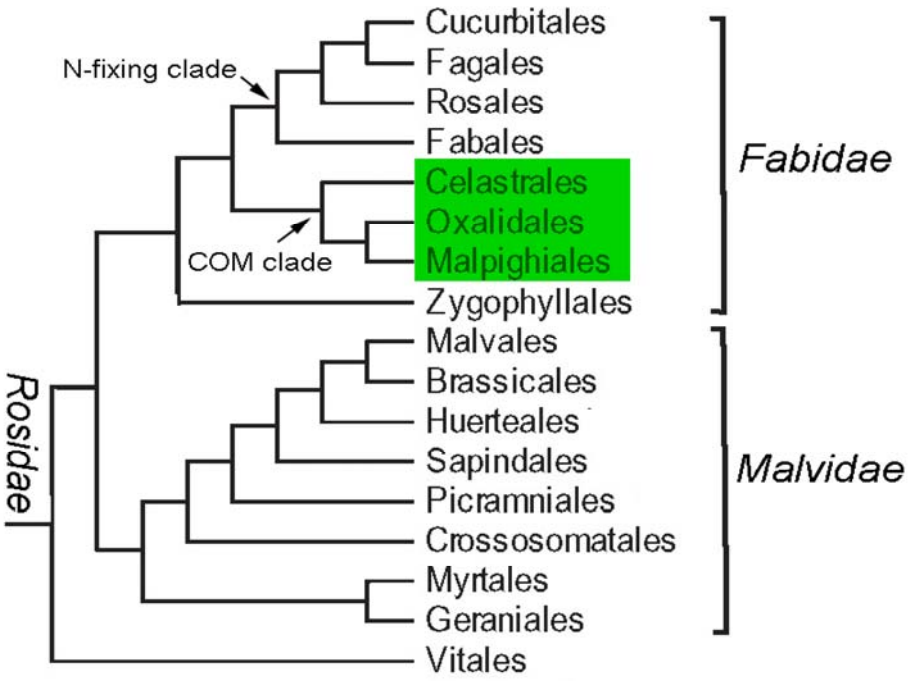


图 1-1 COM 支在薔薇类中的系统位置（APG III 2009）

Figure 1-1 Phylogenetic placement of the COM clade in *Rosidae* (APG III 2009)

在以叶绿体基因为主所建立的被子植物系统发育框架中，薔薇类包含了基部类群、豆类（*Fabidae*）和锦葵类（*Malvidae*），而 COM 支就是构成 *Fabidae* 的其中一支，与另一固氮支（N-fixing clade）成姊妹关系（图 1-1; APG III 2009; Ruhfel et al. 2014）。然而，线粒体、核基因的分子系统学研究结果以及形态结构的研究结果一致支持 COM 支与 *Malvidae* 更近缘（Endress and Matthews，2006；Zhu et al. 2007; Qiu et al. 2010; Zhang et al. 2012）。至此，COM 支系统位置成为薔薇类内部一级分支中系统发育关系不确定而备受关注的类群。

1.1.2 COM 支的研究历史及存在问题

早期的被子植物系统发育研究工作，虽然在取样范围上涉及 COM 支，但由于类群和基因取样的局限，COM 支并没有得到清晰的分辨（图 1-2-1；Chase et al. 1993; Källersjö et al. 1998; APG 1998）。Savolainen et al.（2000a）利用 589 条叶绿体单基因 *rbcL* 且取样上几乎涵盖了真双子植物所有科的系统发育研究中，也仅从拓扑结构上识别了 COM 支及其成员，并表明 COM 支与 *Fabidae* 近缘，但 BS（Bootstrap）支持率不足 50%。同年，Savolainen et al.（2000b）在叶绿体基因

rbcL 基础上增加了 *atpB* 基因，对 357 个有花植物类群进行了系统发育重建，此二基因无论是单独或联合分析均显示了类似上述的 COM 支格局，BS 支持率依然不足 50%。此后，随着取样物种数目和基因数目的增加，Soltis et al.（2000）联合 18S rDNA，*rbcL* 和 *atpB* 基因用最大简约法对 560 种被子植物进行了系统发育重建，在此分析中，COM 支的单系性仅得到 51% JK（Jackknife）支持，而其与 *Fabidae* 的近缘关系得到了 77% JK 支持。就 COM 支的系统位置而言，APG II（2003）的研究结论与上述分析结果并无异同（图 1-2-2）。Hilu et al.（2003）在基于叶绿体基因 *matK* 的被子植物系统发育重建中，COM 支的单系性得到 60% JK 支持，与 *Fabidae* 的聚类关系分别得到 52% JK 与 1.0 的 PP（Posterior probabilities）支持。此外，Soltis 实验室的另外两个类群取样等同、算法不同、且基因取样逐渐增加的系统发育研究结果一致支持 COM 支聚在 *Fabidae* 内，且 BS 支持率高达 89%（Soltis et al. 2007; Burleigh et al. 2009；表 1-1）。而在 APG III（2009）中，COM 支包含类群的范围更大，各目之间的关系趋稳定。此时，COM 支与 *Fabidae* 内的固氮支成姊妹关系，而后 *Fabidae* 与 *Malvidae* 再呈姊妹关系，一起构成蔷薇类的两大亚支（图 1-1）。此结果与 Wang et al.（2009）利用 12 个基因（2 核基因和 10 个叶绿体基因）117 个类群的分析结果基本相同，但后者 BS 支持率达到 100%（表 1-1）。Jansen et al.（2007）和 More et al.（2010, 2011）利用叶绿体基因组数据的分析中，COM 支的位置未有大的变化，支持率却各有差别；而 Soltis et al.（2011）联合叶绿体、线粒体、核三个基因组的 17 个基因联合分析中，COM 支与 *Fabidae* 近缘关系的 BS 支持率却显著降低（57%；表 1-1）。

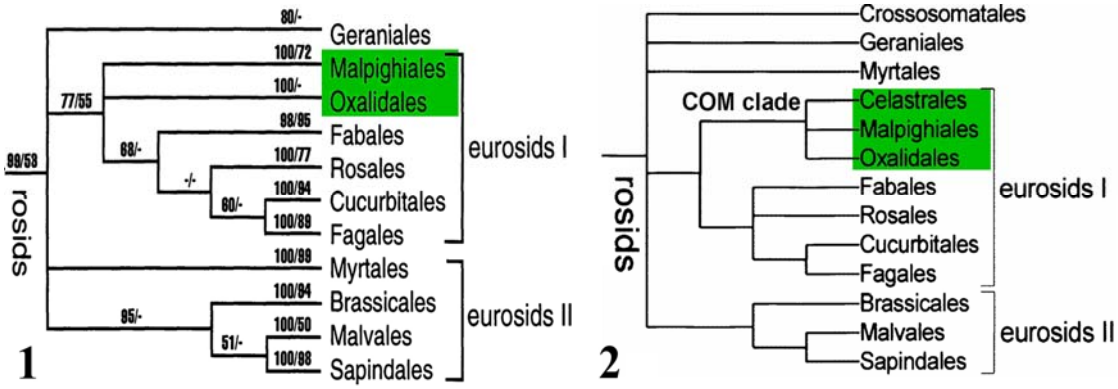


图 1-2 早期 COM 支的系统位置（1：APG 1998；2：APG II 2003）

Figure 1-2 Phylogenetic placement of COM clade in early studies (1: APG 1998; 2: APG II 2003)

由此可见，叶绿体基因组的数据单独（如，Hilu et al. 2003; Moore et al. 2010 2011），或与核基因联合（如，Soltis et al. 2000, 2004; Wang et al. 2009），或叶绿体、线粒体、核的三基因组的联合（Soltis et al. 2011）分析都代表了以叶绿体基因为主导的结论，一致支持 COM 支与 *Fabidae* 的聚类关系（图 1-1），但支持并不稳定（表 1-1）。

Zhu et al.（2007）在基于线粒体 *matR* 基因的蔷薇类系统发育研究中，不支持以往的 COM 支与固氮支的姊妹关系，提出 COM 支应从 *Fabidae* 中分离出来，而与 *Malvidae* 成姊妹关系（图 1-3-1），但此结论仅得到 54% BS 支持。Qiu et al.（2010）利用四个线粒体基因（*atp1*，*matR*，*nad5* 和 *rps3*）对被子植物系统发育关系重建中，进一步加强了 Zhu et al.（2007）的观点，且 BS 支持率达到了 99%（图 1-3-1）。尽管目前基于线粒体在大、中尺度上的系统发育研究比较少，但是线粒体关于 COM 支系统发育位置的结论却与形态性状的研究吻合。Endress 和 Matthews（2006）从花形态结构角度首次提出原被认为属于 *Fabidae* 的 COM 支与 *Malvidae* 有着潜在姊妹群关系，不支持原豆类的单系性，然而这在此前的分子研究中并没有发现。此外，他们从花结构特征上还发现所取样的 COM 支 22 个科的类群与 *Malvidae* 中 18 个科的类群具有一个最显著的共近裔性状—“珠被的内层在受精时比外层更厚”（图 1-4）。还有一些特征在 *Malvidae* 和金虎尾目中比较常见，而在卫矛目和酢浆草目却不普遍，如花瓣扭转状、多雄蕊化、多心皮化，以及珠被与珠被、珠被与珠心之间呈现出彼此分离的趋势。此后，Endress et al.（2013）在金虎尾目的花形态结构的研究中再次认为 COM 支与 *Malvidae* 的姊妹群关系能够更真实地反映蔷薇类内次级分支的系统发育关系。

核基因由于其自身的基因重复与丢失、多拷贝等属性，在被子植物系统发育研究工作中未能得到广泛的开发和应用。先前，被子植物大尺度系统发育分析选用最多的核核糖体 18S rDNA 和 26S rDNA，但往往单独分析分辨不足或与多数的叶绿体基因联合分析而导致核基因结论的独立性不足（表 1-1; Burleigh et al. 2009; Wang et al. 2009; Soltis et al. 1997, 2000, 2007, 2011）。尽管取样没有完全涵盖 COM 支类群，但 Lee et al.（2011）利用核基因组数据的种子植物系统发育分析也显示了 COM 支与 *Malvidae* 的近缘关系（表 1-1）。Zhang et al.（2012）利用五个单拷贝核基因（*SMC1*、*SMC2*、*MCM5*、*MLH1* 和 *MSH1*）对种子植物进行系统发育关系重建时，也得到了同样的 COM 支系统位置，且得到了 BS>95% 高度支持（表 1-1；图 1-3-2）。尽管取样和研究目标不同，另外一些研究也直接

或间接地表明 COM 支与 *Malvidae* 存在更为密切的系统发育关系（表 1-1; Duarte et al. 2010; Finet et al. 2010; Shulaev et al. 2010; Burleigh et al. 2011; Morton 2011）。

显然，以往基于叶绿体基因的大多数研究都认为 COM 支与 *Fabidae* 近缘，而基于线粒体、核基因的系统发育研究则认为 COM 支与 *Malvidae* 更近缘，且得到形态性状的支持。

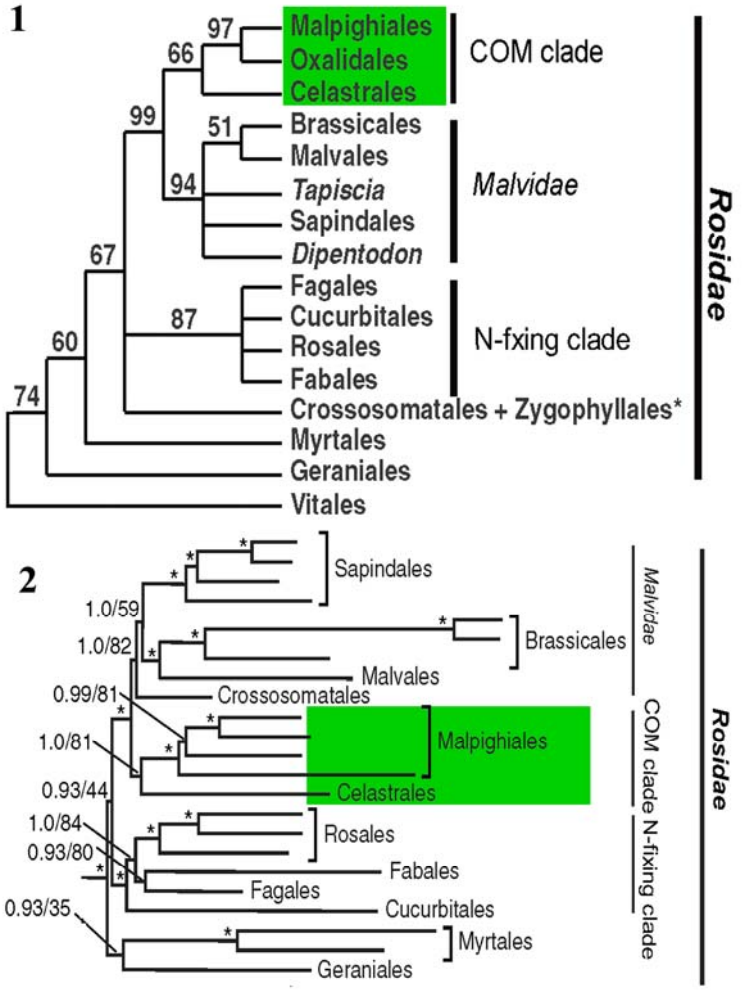


图 1-3 线粒体与核基因系统发育分析中 COM 支的系统位置

Figure 1-3 Phylogenetic placements of the COM clade inferred from mitochondrial and nuclear genes

注：1：线粒体基因系统发育分析支持 COM 支与 *Malvidae* 成姊妹关系（Qiu et al. 2010）；2：核基因系统发育分析支持 COM 支与 *Malvidae* 成姊妹关系（Zhang et al. 2012）。

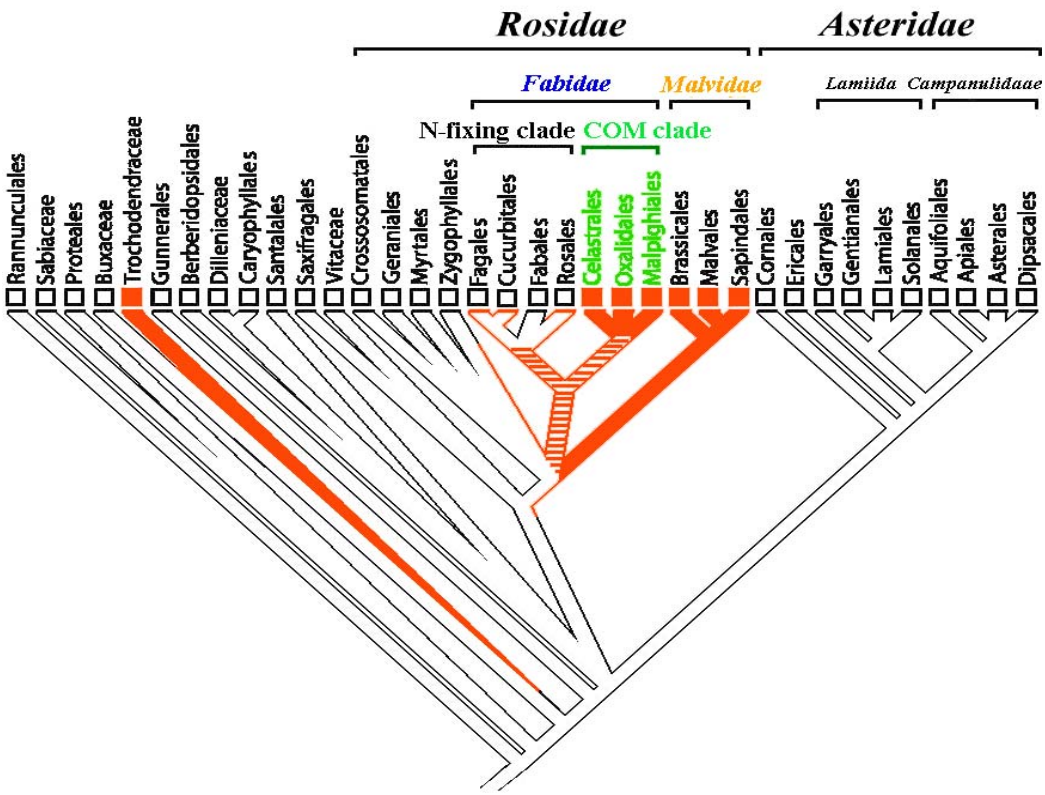


图 1-4 COM 支与 *Malvidae* 花结构上的形态共近裔特征（Endress and Matthews 2006）

Figure 1-4 Morphological synapomorphous shared between the COM clade and *Malvidae* in floral structure (Endress and Matthews 2006)

注：棕色区域代表“珠被的内层在受精时比外层更厚”的共衍征。

1.2 系统发育关系冲突的研究概述

随着系统发育研究的不断发展，人们已经意识到仅依靠单基因或单一来源的数据重建的系统发育关系可靠性不足，甚至还有可能被误导（Wendel and Doyle 1998; Philippe et al. 2005; Salichos and Rokas 2013）。因此，多基因、基因组以及其他源数据的系统发育重建受到系统学家们的一致推荐（Delsuc et al. 2005; Lee et al. 2011; Soltis et al. 2011; Ruhfel et al. 2014）。与此同时，随着下一代测序技术（如，Roche 454, Illumina Solexa, ABI SOLiD）繁荣发展以及系统发育基因组时代的到来，海量转录组、基因组的数据成倍俱增（Soltis et al. 2009b; Soltis and Soltis 2013）。系统发育基因组学的在重建生命之树澄清各大支及深层的系统发

育关系的作用和威力日益显现，但同时也面对着巨大的挑战，即大量基因序列数据各自所反映的复杂进化历史之间、及其与物种本身的进化历史间、以及与局限的建树分析方法之间的冲突越来越突出，越来越普遍（Rokas et al. 2003; Rokas and Carroll 2006; Galtier and Daubin 2008）。在一些情况下，将数据异质的或具有不同进化历史的多基因序列联合用于系统发育重建分析，会得到支持率高但错误的系统发育树（Phillips et al. 2004; Delsuc et al. 2005; Penny et al. 2008; Salichos and Rokas 2013）。比如，诸多研究指出“ANITA”（Amborellaceae, Nymphaeaceae, Illiciaceae, Trimeniaceae 和 Austrobaileyaceae 的缩写）为被子植物的基部类群（Qiu et al. 1999; Mathews and Donoghue 1999; Soltis et al. 2004; Stefanovic et al. 2004; Drew et al. 2014），而 Goremykin 等基于叶绿体全基因组序列提出单子叶植物（禾本科为代表物种）为被子植物的最基部类群（Goremykin et al. 2003, 2004, 2009, 2013），并指出 ANITA 位置受到进化速率异质性的影响（Goremykin et al. 2013）。然而 Soltis 夫妇和 Stefanovic 等分别发现在增加关键类群后，ANITA 依然是被子植物的最基部类群（Soltis and Soltis 2004; Stefanovic et al. 2004），此结论在增加取样量后的叶绿体系统发育基因组学研究中，及针对快速变异位点的分析的结果中，也得到了验证（Jansen et al. 2007; Moore et al. 2007; Drew et al. 2014），而 Goremykin 等的结论很大程度上是由于取样上仅以长枝的禾本科植物作为单子叶植物的代表，和以买麻藤属（*Gnetum*）植物为外类群，而造成单子叶被吸引到基部的假象（Soltis and Soltis 2004; Stefanovic et al. 2004; Drew et al. 2014）。

1.2.1 基因树和物种树的概念

根据系统发育树所反映的是物种间真实的进化关系还是某个（些）基因的进化历史，又有物种树（species tree）和基因树（gene tree）之分。利用某一基因或若干基因所承载的系统发育信息重建的系统发育树，称之为基因树（gene tree），它代表了这些基因的进化历史；而反应物种之间真实进化关系的系统发育树，称之为物种树（species tree），它代表了物种水平上的进化历史（Maddison 1997）。虽然基因树不等同于物种树，但是基因树是可以看作是各基因进化历史的“马赛克（mosaic）”，它反应物种树。无论是基因树间的冲突还是基因树与物种树间冲突，均表明基因树没有正确反映物种树。尽管物种的进化历史难以再

现，我们很难获取绝对的物种树，但我们可以利用现有的数据，探讨各基因进化历史，寻找冲突的根源，调和基因树与物种树的矛盾，最大限度地缩小基因树与物种树间的差异，重建可靠的系统发育关系。

1.2.2 冲突原因以及解决策略

目前已有不少探讨系统发育关系冲突原因及其解决方法的研究和综述（如，Maddison 1997; Wendel and Doyle 1998; Delsuc et al. 2005; Zou and Ge 2008; Degnan and Rosenberg 2009）。Seelanan et al.（1997）根据冲突的显著与不显著认为造成系统发育关系冲突的原因可分为软冲突（soft incongruence；数据不足、支持率较低等原因）和硬冲突（hard incongruence；生物过程）。Wendel and Doyle（1998）将系统关系冲突归纳为三类：技术原因，生物过程以及基因和基因组进化过程（gene and genome-level processes）。Rokas et al.（2003）认为分析因素（analytical factors；数据不足、取样偏差和模型错配等）和生物因素（biological factors；自然选择或基因漂变的作用）是造成系统发育关系冲突原因，而 Galtier and Daubin（2008）将造成不同基因树之间存在冲突的主要原因归纳为人为原因（artefactual reasons；建树错误、随机误差、系统误差等）和生物原因（biological reasons；不完全谱系筛选、旁系同源 paralogous 和基因水平转移）。鉴于上述关于系统发育关系冲突的归纳和总结各有侧重，我们综合考虑引起冲突的原因、能否产生正确的基因树以及对应解决策略等方面的因素，沿用 Seelanan et al.（1997）的概念，但将其发展并总结为新的软冲突（或假冲突）和硬冲突（表 1-2）。软冲突通常是指由于各种分析方法或实验设计（如取样不足）的缺陷所引起的，导致数据内部存在异质性、系统发育信息存在大量非同源相似（homoplasy）等系统发育“噪音”而得出错误的、不稳定的或者支持率低的拓扑关系，即错误的基因树。反之，如果能够得出正确的基因树，但是基因树彼此不一致，或与物种树也存在冲突，那么这种冲突类型定义为硬冲突。

一、软冲突主要包括：人为因素和序列因素。

（一）、人为因素，包括数据不足、取样偏差、基因选择不当、测序错误等（Seelanan et al. 1997; Wendel and Doyle 1998; Delsuc et al. 2005）。显然，分子系统发育研究早期出现的一些基因树冲突往往是由于测序技术的限制使得所用基因片段较短或信息量不足而产生的，而这种冲突现象会因取样偏差会更突出。

但是，下一代测序技术与基因组时代的到来，基因片段的长度与信息量不再成为问题。因此，增加基因信息位点的容量是降低上述随机误差的有效方法（Delsuc et al. 2005; Wortley et al., 2005; Jian et al., 2008）。

表 1-2 导致系统发育关系冲突的原因与解决方案

Table 1-2 Causes and resolutions to phylogenetic incongruence

冲突原因	原因阐述		解决方案
软冲突	人为因素	数据不足	下一代测序技术 系统发育基因组学 取样的代表性与合理密度 基因选择 合理的进化模型
		取样偏差	
		基因选择不当	
		测序错误	
	序列因素	碱基替代模式偏差	第三密码子排除 氨基酸序列建树 RY 编码 快速进化位点移除 一致网络分析法
		长枝吸引	
		进化速率异质性	
		进化饱和	
硬冲突	生物过程	快速辐射分化	最小遗传距离法 融合法 基因树简约法 网状进化网络分析法
		杂交/渐渗	
		不完全谱系筛选	
		基因水平转移	
		旁系同源	
		基因重复与丢失	
		基因重组	

在选择取样时，尽可能地利用现行的系统发育信息和类群分类学知识来有代表性地覆盖类群的多样性。被子植物的各属的物种从单种属到多达 1000 种的大属而不等。生物多样性主流研究显示，在众多属中，一个属一般包含 10 或者更少的种，多达 300 种的大属并不常见（Dial and Marzluff 1989; Scotland and Sanderson 2004）。因此，建议如下取样原则：当属内种数在 1–25 时，可选取 1 个代表物种来取样；当 25–100 时，2 个物种代表；当 101–250 时，3 种代表，

以此类推。但如果研究类群的尺度较小，可根据当前分类阶元之下各亚阶元依据分类学代表性（或者地理分布）适当地增加取样。

同样，基因的选择及其进化速率是重建可靠系统发育关系的关键因素。如果所选基因的进化速率太慢，所能提供的系统发育信息不足，那么系统发育关系将不能很好的解决；如果所选基因的进化速率太快，那么正确的系统发育信息又会被大量的非同源相似信号给淹没，会导致长枝吸引（Felsenstein 1978）。总的来说，如何选择合适的基因或 DNA 片段进行系统发育关系重建一直是个颇具争议的问题（de Queiroz et al. 1995; Wendel and Doyle 1998），也没有很好地能够推选出有效信息位点基因的参数（Rokas et al. 2003）。Rokas et al.（2003）通过矩阵的模拟分析建议至少选用 20 个非连锁的基因才能够得出具有相当高支持率、可靠的物种树，才能避免个别或者少数的基因得出的系统发育关系的风险。一般来说，用于建树的序列应该满足：1)、碱基替换速率能提供足够的系统发育信号；2)、足够长度的核苷酸序列所包含的信息位点足以克服取样错误。Hillis（1996）通过矩阵模拟探讨核苷酸的数量与重建系统发育准确性的关系，他指出应用与 *rbcL* 相近长度的 5000 多个核苷酸信息位点的矩阵就能让 90%的建树模型正确地推导出系统发育关系。目前普遍认为，通过增加数据量和应用更合适的分析方法能够消除人为因素对系统发育重建的影响（Rokas et al. 2003; Delsuc et al. 2005; Wortley et al. 2005; Yang and Rannala 2012）。

（二）、序列因素，即序列本身的组成成分、替换饱和以及进化速率的不一致性导致分子数据与建树模型错配得出错误的拓扑结构（Jeffroy et al. 2006）。常见的序列因素包括碱基替代模式偏差（compositional bias）、长枝吸引（long-branch attraction）、碱基位点进化速率异质性（heterotachy）、进化饱和（evolutionary saturation）等（表 1-2；Wendel and Doyle 1998; Delsuc et al. 2005）。

碱基替代模式偏差。不同生物物种间核苷酸组成成份或碱基的替代模式是具有差异的。而目前我们常用的建树模型都基于同一假设，即所有生物类群在进化时都遵循相同的碱基替换模式，当类群因突变发生偏性或者显著异质时，那么相似的核苷酸组成或替代模式，就会导致不相干的类群依据模型中的算法机械地聚在一起，得出支持率高，但是错误的拓扑关系。

长枝吸引。主要是由于谱系间的进化速率高度不一致，一些进化速率较快的类群由于频繁的突变会在许多碱基位点随机地被替换上相同的碱基，这就形成了非同源相似。在建树过程中，非同源相似会使得不是来自于共同祖先的序

列的代表性分类群相互“吸引”聚在一起，造成了长枝吸引，从而掩盖真实的系统发育关系，得到错误的拓扑结构。

碱基位点进化速率异质性。不同碱基位点因选择压力不同而使得基因具有不同的进化速率。随着时间推移，选择压及基因或者蛋白给定位点上的进化速率亦随之变化。然而，目前我们常用的建树模型都假设特定碱基位点的替代速率随着时间推移在类群之间保持恒定，这种假设与实际的分子进化过程并不完全吻合。而相反另一些不相关类群因具有相当比例的恒定位点而有可能聚在一起，这种现象会导致潜藏较深、假的系统发育关系，且极其不容易在序列上检测到（Delsuc et al. 2005）。

另外，基因组尺度上的突变事件—稀有基因组变化（Rare genomic changes, RGCs），如逆转录组整合、编码序列的插入和缺失、内含子的获得与丢失等，尽管这些变化没有独立、精确的发生方式，但也能在一定程度上造成非同源相似，（Delsuc et al. 2005; Rokas and Carroll 2006）。

目前针对于数据异质性和非同源相似等序列因素造成的冲突，除了多基因联合外，有效可行的方法还有第三密码子移除法、RY 编码法、快速进化位点移除法等（表 1-2）。这些方法可以降低核苷酸序列中的置换饱和、进化速率和碱基组成成份的异质性。针对长枝吸引，可以通过增加类群取样和分子性状来打破长枝，从而建立正确的系统发育树（Xiang et al. 2002）。对于编码蛋白的基因，相对于第一、二位密码子，第三位密码子受功能制约较小，进化速率通常较快。因此通过排除第三密码子，有时是能够建立正确的系统发育树的简单有效方法。同时，受基因功能和选择的制约，碱基间发生转换的频率要高于颠换，因此将四种核苷酸状态，归为 purines（A, G = R）和 pyrimidines（C, T = Y）两类，即 RY 编码法，也能够有效减少置换饱和给系统发育分析带来的噪音（Delsuc et al. 2003; Phillips and Penny 2003; Phillips et al. 2004; Gibson et al. 2005）。移除快速进化位点也是排除快速位点给系统发育关系带来误导的另一种行之有效的方法（Brinkmann and Philippe 1999; Lopez et al. 2002; Burleigh and Mathews 2004; Pisani 2004; Phillips et al. 2005）。该方法是先对矩阵所有位点的进化速率进行计算、排序，然后按照进化速率快慢的顺序逐渐移除快速进化的位点，利用剩余位点反复建树，直至拓扑结构不再发生变化，从而得出稳定而可靠的系统发育关系（Goremykin et al. 2010）。

此外，还有系统发育网络分析法（phylogenetic networks）概念下的分支网

络分析法（splits networks；表 1-2），它可以反映数据中冲突、系统位置不稳定等与一致性不兼容的信号（Huson and Bryant 2006）。其中，一致网络分析法（consensus networks）比较常用（Holland et al. 2004, 2006; Huson and Bryant 2006; Zou et al. 2008）。其本质是一个在最大似然或最大简约的原则下参与系统发育分析所有基因的基因树的统计集合，基因树的分支模式与网络中类群的分支格局（网络的边分支方式）相对应。它体现所有基因树传递的系统发育关系，包括冲突的、不稳定的分支结构（Holland and Moulton 2003; Holland et al. 2004; Huson and Bryant 2006）。根据其分支某个节点分歧可能性的基数会出现一个对应基数维的超立方体以表示该节点有多少种拓扑结构被一定比例的基因树支持；网络中边的长度代表支持该分支的支持率或者与该分支模式一致的基因树的比例（或遗传距离）。一致网络分析法优点在于它比单纯地将多基因数据联合给出一个最优拓扑结构的传统做法更具有信息，它更能展现参与分析基因支持的所有拓扑关系（包括软冲突和硬冲突中的因素），它能把数据内潜藏的不确定性可视化（Holland and Moulton 2003; Holland et al. 2004）。常见的用于该分析的软件为 SplitsTree4（http://ab.inf.uni-tuebingen.de/software/splitstree4/welcome.html）。

另外，如果数据与模型不匹配的话，会产生与真实系统发育信号相抗衡的错误信号。这些噪音通常会在序列中随机分布，现行的建树方法可以从中提取更多的信息位点。但是，如果生物进化信息位点的信号很微弱，那么重建系统发育关系时，噪音就会占主导。这样的话，也会产出错误的系统发育关系。故加大基因数据量以及数据来源，熟悉掌握各建树模型的优点与不足，有针对性地将不同的建树方法和模型用于系统发育重建（Yang and Rannala 2012）。

二、硬冲突主要指不同的基因可能揭示或反映不同的进化过程，是具有遗传基础的，包括快速辐射分化（rapid radiation）、杂交 / 渐渗（hybridization/introgression）、不完全谱系筛选、基因水平转移、旁系同源、基因重复和丢失和基因重组等（表 1-2；Doyle 1992; Maddison 1997; Linder and Rieseberg 2004; Galtier and Daubin 2008）。物种在相对短的分子进化尺度内快速分化，那么通过此类群的系统发育重建就会得到内部分支较短且置信度较低的基因树。因此，快速物种分化造成较短的内部分支也是误导系统发育的因素之一（Wendel and Doyle 1998）。越近缘或分化时间越短的物种间（快速辐射分化的类群间），发生杂交或基因渐渗事件就会越频繁；杂交其本质上是基因渐渗的延伸（Simpson 2012）。双亲遗传的核基因树与细胞质单亲遗传的叶绿体（或线

粒体）基因树之间冲突往往是物种间杂交事件造成的格局。Rieseberg 和 Soltis（1991）已经指出杂交/渐渗事件在植物中很普遍。此后，在被子植物中就有叶绿体基因渐渗现象报道 100 多例（Rieseberg et al. 1997）。谱系筛选一般发生在祖先类群在很短时间内经历连续物种分化事件的类群中，就是说，具多态性祖先还没有足够的时间持续分化为该谱系的单系类群，第二次物种分化事件就开始了，此时极可能产生该谱系的基因树与物种树冲突。特别是分化时间间隔越短，有效种群越大，不完全谱系筛选越可能发生（Pamilo and Nei 1998; Rokas and Carroll 2006; Galtier and Daubin 2008; Degnan and Rosenberg 2009）。基因水平转移是相隔较远的物种间的非有性生殖的基因交流，尤其是寄生类群之间（Davis and Wurdack 2004）。其冲突效果与杂交/渐渗事件类似，所以各个基因所得的基因树会因基因间交流的次数和性质不同而不同（Galtier and Daubin 2008）。类似地，基因由于基因重复事件产生了两或多个拷贝；进化过程中某些拷贝在部分类群中丢失，导致不完全的拷贝在后代中保留下来，那么利用这个基因建树就会得到与物种本身的进化历史不一致的基因树（Page and Charleston 1997）。基因重组通常发生在物种各世代的减数分裂期间的核基因组内，若干世代之后，该基因组内就会富集各种复杂的进化历史（Wang et al. 2002）。而且在重组的过程中常常伴随着基因漂变和自然选择，从而导致不同的谱系遗传定位到等位基因的不同位点上，因此造成的冲突效果与不完全谱系筛选类似（Linder and Rieseberg 2004）。但叶绿体和线粒体属于细胞器单亲遗传，通常它们不经历有性生殖或者减数分裂的基因重组（Linde and Rieseberg 2004）。

上述各种原因给重建可靠的生命之树带来了很大挑战，如果不深入细致地探讨基因或者基因组间冲突的原因，而盲目地进行多基因联合分析，有时产生的不可信的、甚至是错误的拓扑关系，更不能真实地反应物种关系。事实上，硬冲突（生物过程）更具有重要的进化意义，我们需要分析影响拓扑结构变化的数据源，分析其可能发生的生物过程来调节并且解释基因树与物种树之间的冲突（Page and Charleston 1997）。硬冲突一般出现在来自不同基因组的基因之间，或者非连锁的核基因之间。基因水平转移比较容易鉴别，通常发生在关系非常远的类群之间，尤其是寄生类群之间（Davis and Wurdack 2004）。可杂交和不完全谱系筛选造成的冲突很容易通过网状进化网络展现出来，但二者均能产生相似的拓扑结构而难以相区别。为此近年来提出了如下几种统计分析方法。根据谱系筛选发生在物种分化事件之前，而杂交发生在物种分化事件之前或者之后，

Joly et al.（2009）提出了基于统计参数的最小遗传距离法（minimum genetic distance method）。融合理论（coalescent theory）认为祖先的多态性经历 5×Ne 代（Ne 是有效群体大小）以后就融合在一起，从而得到一致的基因树（Rosenberg 2003; Degnan and Rosenberg 2009; Rosenberg and Degnana 2010）。Pelser et al.（2010）根据此融合理论提出了融合法（Coalescence-based methods）。Burleigh et al.（2011）应用基因树简约法整合了核基因的重复和丢失事件，基于包含 510,922 条蛋白序列的核基因组数据重建了 136 个植物类群的 18,896 棵核基因树。该研究对于调和物种树与基因树间的冲突具有一定指导和参考作用。检测基因重组事件是种群遗传研究很重要的一个环节，是重建祖先基因组和后代谱系关系的核心（Posada and Crandall 2001; Zhang et al. 2002; Wall and Pritchard 2003）。

通过基因树简约法法和网状进化网络分析法（reticulation networks）是目前解决硬冲突、探讨生物进化过程常用的方法（Huson and Bryant 2006; Burleigh et al. 2011）。近来算法和软件的更新使基因树简约法的搜树速度大大提升，此方法是首次针对基因组数据进行运算（Bansal et al. 2007; Wehe et al. 2008）。基因树简约法就是根据导致基因树冲突的生物事件（基因重复和丢失、不完全谱系筛选、基因水平转移等）发生最少的次数来推测物种树（Goodman et al. 1979; Maddison 1997）。如，调和在基因重复和丢失的进化事件下造成的系统发育关系冲突，就需要在所有基因树的集合中寻找一棵即能代表所有取样的物种还能指示基因重复和丢失事件发生次数最少的基因树来代表物种树（Page and Charleston 1997; Slowinski and Page 1999; Page 2000）。系统发育网络分析法中的一致网络分析法和网状进化网络分析法都可分析、展现生物网状进化造成的拓扑结构分歧、不一致的系统发育现象，比基因树简约法更直观、更具有信息。网状进化网络分析法主要是针对生物过程（硬冲突）的分析，如杂交、基因水平转移等，其内部节点代表祖先类群，分支样式代表网状进化事件（Linde and Rieseberg 2004; Gusfield and Bansal 2005; Holland et al. 2006）。SplitsTree4 也可用于该分析方法。一致网络分析法，能充分展现参与分析基因支持的所有拓扑关系，可将数据内潜藏的不确定性可视化（Holland et al. 2004），它能暗示进化历史，但并不能代表物种所经历过；内部节点也不代表祖先类群（Holland et al. 2006）。

随着测序技术的日趋繁荣，测序成本极大地降低（Soltis et al. 2009a, 2010; Hamilton and Buell 2012）。据统计，从 1982 年距今，在 GenBank 的数据库中存储了近 2 亿条 DNA 序列，而这些序列每 18 个月内就会翻倍（NCBI Website:

http://www.ncbi.nlm.nih.gov/genbank/statistics）。基于如此海量的数据，在重建生命之树的过程中，产生系统发育关系的软冲突和硬冲突是不可避免的。被子植物的遗传背景复杂，因此，有时利用单一基因、多基因和基因组数据，不能完全揭示被子植物的进化历史；有时应用了整个叶绿体基因组数据，被子植物某些关键类群的系统位置仍然无法确定（Moore et al. 2010）。这就要求我们广泛地利用多源生物信息数据，不局限于单亲遗传叶绿体和线粒体细胞器基因，更要关注蕴含丰富系统发育信息的、双亲遗传的、直系同源的核基因；而且，核基因还有助于研究和发现被子植物进化过程中的不完全谱系筛选、杂交/渐渗等现象（Zeng et al. 2014）。此外，还需要结合形态、化石等非分子数据来佐证、支持系统发育关系。唯此才可以进一步验证前人的研究成果以及解决生命之树疑难的分支节点，并揭示潜在系统发育关系冲突背后的生物进化事件，最终以共同联合、相互验证、追根溯源的模式来重建准确的、可靠的生命之树。

1.3 研究目的意义

以上我们综述了 COM 支的研究历史和造成系统发育关系冲突的各种因素，了解到线粒体、核基因以及形态研究结果都支持 COM 支与 *Malvidae* 近缘，这与叶绿体基因支持 COM 支和 *Fabidae* 近缘的结论相冲突（表 1-1），又考虑前人研究的取样策略、建树方法、以及选用的基因也不尽相同（表 1-1），所以 COM 支系统发育关系的冲突可能来自软冲突（取样、序列者建树方法中的系统误差），也可能是硬冲突（不完全谱系筛选、渐渗杂交等生物过程）。因此，围绕 COM 支的系统位置，我们提出如下两个科学问题：

1）、COM 支系统位置的冲突是否真实存在？如果存在，这种冲突是系统误差或取样偏差造成的还是生物过程造成的？

2）、如何利用核基因组数据理解 COM 支的系统位置冲突？

基于上述科学问题，本研究设计了代表植物基因组（叶绿体、线粒体、核基因组）的、类群取样能够代表被子植物系统发育框架各大分支的三个矩阵；然后运用多种性状编码和数据筛除的分析方法来验证 COM 支的冲突的系统发育关系是否由系统误差所致。另外，我们还设计了单拷贝（single-copy）和多拷贝（multi-copy）核基因组矩阵，进一步探讨了 COM 支系统位置是否与上面提到的硬冲突中的不完全谱系筛选或渐渗杂交生物进化事件有关。具体研究分析

流程见图 1-5。

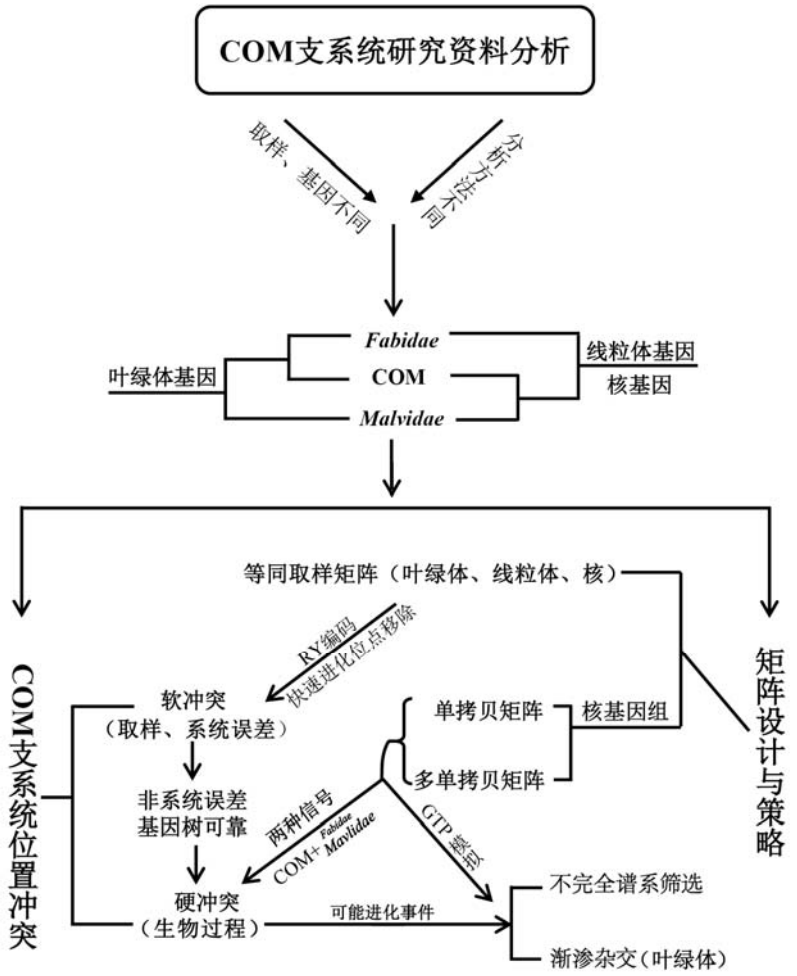


图 1-5 COM 支系统发育研究分析流程

Figure 1-5 The analytical framework in this study on COM clade

本研究围绕目前生命之树分支疑难节点上争论的系统发育关系的焦点展开，旨在解决蔷薇类一级分类上的重大分歧，为构建可靠的蔷薇类系统发育框架奠定基础，COM 支系统位置冲突问题的解决对于理解蔷薇类，以至整个被子植物的系统发育与进化都是极为关键的。本研究不仅为验证、解决不同基因组间物种类群系统发育关系冲突的问题提供了借鉴和参考作用，而且还展示出了基因组数据在揭示生命之树深层次系统发育关系的分歧、复杂生物过程方面的重要意义。

此外，COM 支类群（如，卫矛科 Celastraceae、酢浆草科 Oxalidaceae、大戟科 Euphorbiaceae 等植物）具有较高的经济、药用价值和园林观赏价值。卫矛科一些类群在全世界范围内作为抗癌药物的研究对象，已引起了人们极大的关注；卫矛科的巧茶属（*Catha*）植物在非洲和阿拉伯半岛作为兴奋剂和抗疲劳药用植物被广泛栽培（Simmons et al. 2008）。COM 支的物种还包含了丰富的形态和生态多样性，如：叶状体半沉水植物（川苔草科 Podostemaceae），全寄生植物（大花草科 Rafflesiaceae），风媒类群（杨柳科 Salicaceae），仙人掌状无叶肉质植物（大戟科）（Wurdack and Davis 2009）。故搞清楚 COM 支类群的系统发育关系，也可以为合理保护与有效地利用这些植物资源提供理论基础。

第二章 材料与方法

为了方便分析和讨论 COM 支的系统位置，在本研究中，我们将蔷薇类内的 COM 支、*Fabidae*、*Malvidae* 分别作为三个不同的类群来处理，尽管在现行的分子系统学分类系统中，COM 支被认为是 *Fabidae*（或 fabids）内的一个亚分支（Cantino et al. 2007; APG III 2009）。

同样地，为了方便后面章节的讨论，我们把目前理论上 COM 支的三种系统位置归纳为：一，代表叶绿体基因数据所支持的 COM 支聚于 *Fabidae* 内聚类关系（以下简称 COM + *Fabidae*；图 2-1-1）；二，代表线粒体与核基因，还有形态性状分析结论所支持得 COM 支与 *Malvidae* 的聚类关系（以下简称 COM + *Malvidae*；图 2-1-2）；三、理论上的第三种可能，即 *Malvidae* 与 *Fabidae* 成姊妹关系，而 COM 支再聚于二者之外的拓扑关系（以下简称 *Fabidae* + *Malvidae*；图 2-1-3）。

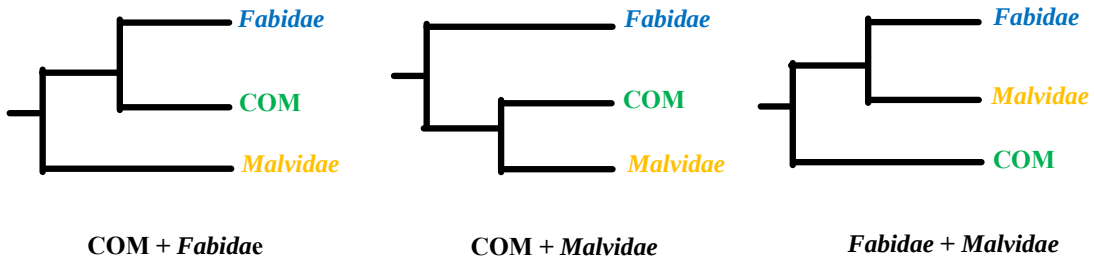


图 2-1 COM 支三种可能的拓扑关系

Figure 2-1 Three hypothetical placements for the COM clade

2.1 取样策略和矩阵装配

分子系统学研究表明，不同的取样策略、不同的取样范围以及选用的基因不同，都会对重建系统发育关系的拓扑结构和分辨率产生影响。鉴于前人研究的取样策略、建树方法、以及所采用基因数据也不尽相同（表 1-1），因此，为了真实反映 COM 支的系统位置在不同基因组间冲突问题，我们从已公开发表的数据库中抽取并装配了分类学上取样相当、且代表植物基因组的三个矩阵：叶

绿体 82 类群 78 基因矩阵（Ruhfel et al. 2014）、线粒体 79 类群 4 基因矩阵（Qiu et al. 2010）以及核 92 类群 5 基因矩阵（Zhang et al. 2012）。

基于三个矩阵源自的研究都是在大尺度上针对整个被子植物的系统发育关系的工作，故其取样策略本身已经代表性地覆盖了包括蔷薇类和 COM 支在内被子植物各大主要分支。如，依据 APG III（2009）系统，Zhang et al.（2012）取样的 92 个类群来自 75 科 46 目覆盖了被子植物 73%（46/63）的目。本研究中的取样类群所在各大分支的目和科的名称与范围均依据 APG III（2009），各大分支的名称依据 Cantino et al.（2007）。在矩阵的装配过程中，我们最大限度地满足三个基因组的数据来自同一科、同一属或同一种；同时也最大限度地匹配取样的系统发育代表性，备选取样类群尽可能来自同一科或目，而对于 COM 支以外的类群则保证至少在同一系统发育分支上。

考虑到叶绿体和线粒体通常基因连锁且单亲遗传，故有必要再增加双亲遗传的非连锁的核基因，来加强对细胞器基因冲突的观察和判别能力。另外，我们还想知道核基因组内 COM 支的系统位置如何。基于这样的认识，本研究装配了两个核基因组矩阵：8,445 个单拷贝核基因直系同源矩阵（orthologs sets）和 3,748 个多拷贝核基因矩阵。其中，单拷贝核基因矩阵来自 Lee et al.（2011）的种子植物的系统发育基因组学研究（BIGPLANT 网站：http://nybg.bio.nyu.edu/），它是目前规模最大的直系同源核基因组矩阵。尽管此矩阵仅包括了 COM 支金虎尾目的 7 个属（木榄属 *Bruguiera*，大戟属 *Euphorbia*，橡胶树属 *Hevea*，木薯属 *Manihot*，杨属 *Populus* 和蓖麻属 *Ricinus*），未能包括卫矛目和酢浆草目，但它是目前能够为 COM 支的系统位置分析提供独立的核基因位点的最佳数据。多拷贝核基因矩阵是从 OrthMCL（Chen et al. 2006）数据库中提取的 22 个陆地植物类群的全基因组数据。22 个陆地植物类群包括卷柏（*Selaginella moellendorffii*），小立碗藓（*Physcomitrella patens*）和另外 20 个被子植物（其中，毛果杨 *Populus trichocarpa* 代表 COM 支；野草莓 *Fragaria vesca*，蒺藜苜蓿 *Medicago trunculata* 和野大豆 *Glycine max* 代表 *Fabidae*；拟南芥 *Arabidopsis thaliana*，条叶盐芥 *Thellungiella parvula*，番木瓜 *Carica papaya* 和可可 *Theobroma cacao* 代表 *Malvidae*），从而装配了 3748 个多拷贝核基因矩阵。尽管取样稀疏，但是来自全基因组的序列的信息位点比不完整的转录组数据更能够准确评估诸如基因丢失等生物进化过程。

2.2 序列比对

鉴于取样相当的三个矩阵均来自已发表的数据，故其矩阵本身比对状况已经良好，所以大部分矩阵经过人工检验和编辑后直接使用。仅对因取样设计需要个别调整的类群用 MUSCLE（Edgar 2004）进行自动比对；自动比对后，再用 Geneious（http://www.geneious.com）进行手动校正。

对从 BIGPLANT 网站（http://nybg.bio.nyu.edu/）获得的单拷贝核基因矩阵（Lee et al. 2011），经过人工检验、确认后直接进行使用。

对来自陆地植物全基因组数据库中的 22 个物种 3,748 个多拷贝核基因矩阵使用 MAFFT（Katoh et al. 2005）以默认参数自动比对后，再手动调整使用。

2.3 系统发育分析

2.3.1 核苷酸序列系统发育重建

本研究中对所有矩阵的最大似然法（Maximum Likelihood, ML）分析都是应用 RAxML 7.2.8（Stamatakis 2008）软件在 GTRCAT 模型下进行的。

对三个基因组核苷酸小矩阵，都使用 100 次重复的无参数自展分析来获得最优 ML 树。此外，我们对三个矩阵中各个单基因以同样的方法进行了 ML 分析，并统计和归纳各个单基因中 COM 支系统位置情况。

2.3.2 RY 编码分析

RY 编码被认为可以有效地降低核苷酸序列中的置换饱和以及进化速率和碱基组成成份的异质性，并提高树的可信度（Phillips and Penny 2003; Harrison et al. 2004; Phillips et al. 2004; Delsuc et al. 2005; Gibson et al. 2005）。因此，我们将三个矩阵中的四种核苷酸类型分别编码为 R\Y 两种类型，即将所有的嘌呤编码为 R（A,G = R），将所有嘧啶编码为 Y（C,T = Y），随后对 RY 矩阵以 GTRCAT 模型进行 ML 建树分析。

在进行 RY 编码时，只需要在 nexus 格式文件的 format 命令后添加“equate="G=A T=C"”，模型参数设置为 nst=1。

举例如下：

#NEXUS

Begin data;

Dimensions ntax=10 nchar=1000;

Format equate="G=A T=C" datatype=dna missing=? gap=-;

2.3.3 快速进化位点移除分析

移除核苷酸序列中的快速变异位点可以缓解长枝吸引和模型匹配错误对系统发育分析造成的误导（Brinkmann and Philippe 1999; Philippe et al. 2005; Zhong et al. 2011; Goremykin et al. 2013），此方法已成为重建系统发育关系研究中常用的方法（如，Delsuc et al. 2005; Regier and Zwick 2011; Rajan 2013）。因此，为了进一步验证这些快速变异位点是否与 COM 支系统位置的冲突有关，我们也对三个基因组的矩阵分别进行快速变异位点移除分析。首先，根据 Goremykin et al.（2010）描述的方法将矩阵的进化速率依据观测变异值（observed variability, OV）进行排序；然后对每个排序好的矩阵进行 5%，10%，20%，30%，40%，到 50%的快速位点移除；最后对每次快速位点移除之后的矩阵应用 GTRCAT 模型进行 ML 建树分析。

2.3.4 氨基酸序列分析

用氨基酸序列代替核苷酸序列来建树，能够缓减或避免序列间碱基组成的异质性以及核苷酸序列中碱基位点的转换和颠换带来的非同源相似，使得所得到的系统发育关系更可靠（Loomis and Smith 1990; Hasegawa and Hashimoto 1993; Hashimoto et al. 1995）。故我们还把叶绿体、线粒体以及核基因的核苷酸矩阵转换为氨基酸（Amino Acide, AA）矩阵应用 PROTCATJTT 模型（Jones et al. 1992）以最大似然法进行 ML 分析。

2.3.5 单拷贝核基因分析

Lee et al.（2011）首次利用 22,833 个核基因组的直系同源矩阵来重建 101 属种子植物的系统发育关系。尽管这个矩阵的取样并没有完全覆盖 COM 支的类

群，但它是目前能够为 COM 支的系统位置分析提供独立的单拷贝核基因位点的最佳数据。我们对此矩阵的每一棵基因树都进行搜索和评估以检验其对 COM 支的系统位置是否具有系统发育信息。

首先，我们将 Lee et al.（2011）整个核苷酸矩阵分割为 26,986 个不同的小矩阵，每个矩阵代表一个单拷贝直系同源核基因。之后，我们依据它们对 COM 支系统位置的是否具有系统发育信息而进行鉴定和筛选：这些矩阵至少应包含一个 COM 支的物种，一个来自 *Fabidae* 的物种，一个来自 *Malvidae* 的物种，和一个上述类群之外的物种。最后，我们获得了对 COM 系统位置有系统发育信息的 8,445 个小矩阵（即相当于 8,445 单拷贝直系同源核基因），并对每一个小矩阵都应用 RAxML v.7.2.8 以 GTRCAT 模型进行 100 次重复的最大似然法（ML）运算，并分别统计支持 COM 支三种拓扑结构（图 2-1）的基因树数。上述针对 COM 支的三种系统位置支持率的统计分析都是通过 Perl 语言和 Newick 应用程序来实现的（Junier and Zdobnov 2010）。

2.3.6 多拷贝核基因分析

围绕 COM 支的系统位置，除了对单拷贝核基因组数据分析外，我们还对多拷贝核基因家族的数据开展了基因树简约法分析。与上述单拷贝直系同源核基因组数据不同，多拷贝核基因数据所反映的系统发育关系不是那么的直接、明朗。比如，来自 COM 支某一物种的多拷贝核基因序列中，可能一条序列支持其与 *Fabidae* 聚类，而另一条序列则支持其与 *Malvidae* 聚类。为了解决这一问题，我们针对每一给定的 COM 支系统位置（图 2-1），计算其下每一基因树的调和成本（reconciliation cost）。COM 支每一位置，对应三种不同的调和成本（每一调和成本即为一种可能的进化事件）：1）、发生最少基因重复事件（gene duplications）下的基因树数目；2）、发生最少基因重复和丢失事件（gene duplications + losses）下的基因树的数目；3）、发生最少不完全谱系筛选事件（deep coalescence）下的基因树的数目（Maddison 1997）。我们用基因树简约法来鉴定每一基因家族对 COM 支三种不同系统位置的支持，即包含最少进化事件（即，最低调和成本）所对应的拓扑结构，为该基因家族所支持的 COM 支系统位置。如果其中两种或三种系统位置具有等量的调和成本（即等量基因树数）支持，则认为该基因家族对 COM 支的系统位置没有信息。

我们从 OrthMCL（Chen et al. 2006）数据库中筛选出 22 个陆地植物类群的全基因组数据，共装配了 3748 个多拷贝核基因矩阵。对于每一个多拷贝基因的矩阵，都使用 RAxML 以 GTRCAT 模型进行 100 次重复的 ML 建树分析。然后，基于上述方法和原理，我们对每一给定的 COM 支系统位置（图 2-1），用 OptRoot 软件（Andre Wehe：http://www.wehe.us/optroot.html.）统计每一基因家族的基因树在三种模拟进化事件下的调和成本。

另外，基因树的置根很大程度上能够影响到调和成本的评估。而且，通常情况下，对于多拷贝基因树置根比较困难。因此，我们用调和成本最低（包含的进化事件最少）原则给每一多拷贝核基因树置根。

第三章 COM 支系统发育关系分析结果

3.1 多基因数据

尽管我们的取样范围涉及整个种子植物，但本研究结果主要针对蔷薇类内 COM 支与其他各成员的系统发育关系展开。我们通过对三个取样相当的叶绿体、线粒体和核基因矩阵的 ML 分析重现了 COM 支冲突的系统位置（图 3-1–图 3-3，图 A1）。

其中，82 类群 78 基因的叶绿体矩阵的系统发育分析，结果与前人基于叶绿体基因分析的结论一致（APG III 2009; Wang et al. 2009; Moore et al. 2010; Soltis et al. 2011; Ruhfel et al. 2014; 图 3-1）。COM 支自身以及 COM 支与 *Fabidae* 聚类关系都得到了 100% BS 支持。但 COM 支的准确系统位置并不确定，仅 52% BS 支持 COM 支与除 *Bulnesia*（维腊木属，Zygophyllaceae，Zygophyllales）以外的 *Fabidae* 类群成姊妹关系，而 *Bulnesia* 再为上述二者成姊妹群（图 3-1）。叶绿体 78 个基因单独分析中，32 个基因对 COM 支的系统位置没有信息位点，46 个基因对 COM 支的系统位置有信息位点（表 3-1；附录 B）：36 个基因支持 COM + *Fabidae* 的聚类关系（其中，仅 5 个基因的 BS 支持率高于 60%）；5 个基因支持 COM + *Malvidae*，BS 支持率都不足 50%；5 个基因支持 *Fabidae* + *Malvidae*（仅一个基因的 BS 支持率大于 50%；附录 B）。RY 编码分析 100% BS 支持 COM + *Fabidae* 的聚类关系（图 A2-1）。而在氨基酸矩阵建树分析中，COM 支与除 *Bulnesia* 以外的 *Fabidae* 类群的成姊妹关系（BS 支持率仅为 47%），而且以弱支持率将 *Bulnesia* 置于 *Malvidae* 中（图 A2-1）。在快速进化位点移除分析中，随

表 3-1 叶绿体 78 个单基因分析结果

Table 3-1 Results from 78 chloroplast individual genes analysis

COM 支系统位置	**基因数量**	**注解**
COM + *Fabidae*	36	BS > 60% 5 个
COM + *Malvidae*	5	BS > 50% 0 个
Fabidae + *Malvidae*	5	BS > 50% 1 个（81%）
无信息	32	—

着快速进化位点的逐渐移除，COM + *Fabidae* 的拓扑关系很快被打破。如，分别移除快速进化位点的 5%，10%至 20%，那么 COM + *Fabidae* 的支持率会从 98%降到 48%直至 1%。如果移除更多的快速进化位点，则对 COM + *Fabidae* 的拓扑关系没有支持。总之，在叶绿体矩阵的快速进化位点移除分析中，没有发现任何信息位点支持 COM + *Malvidae* 或者 *Fabidae* + *Malvidae* 的拓扑关系。

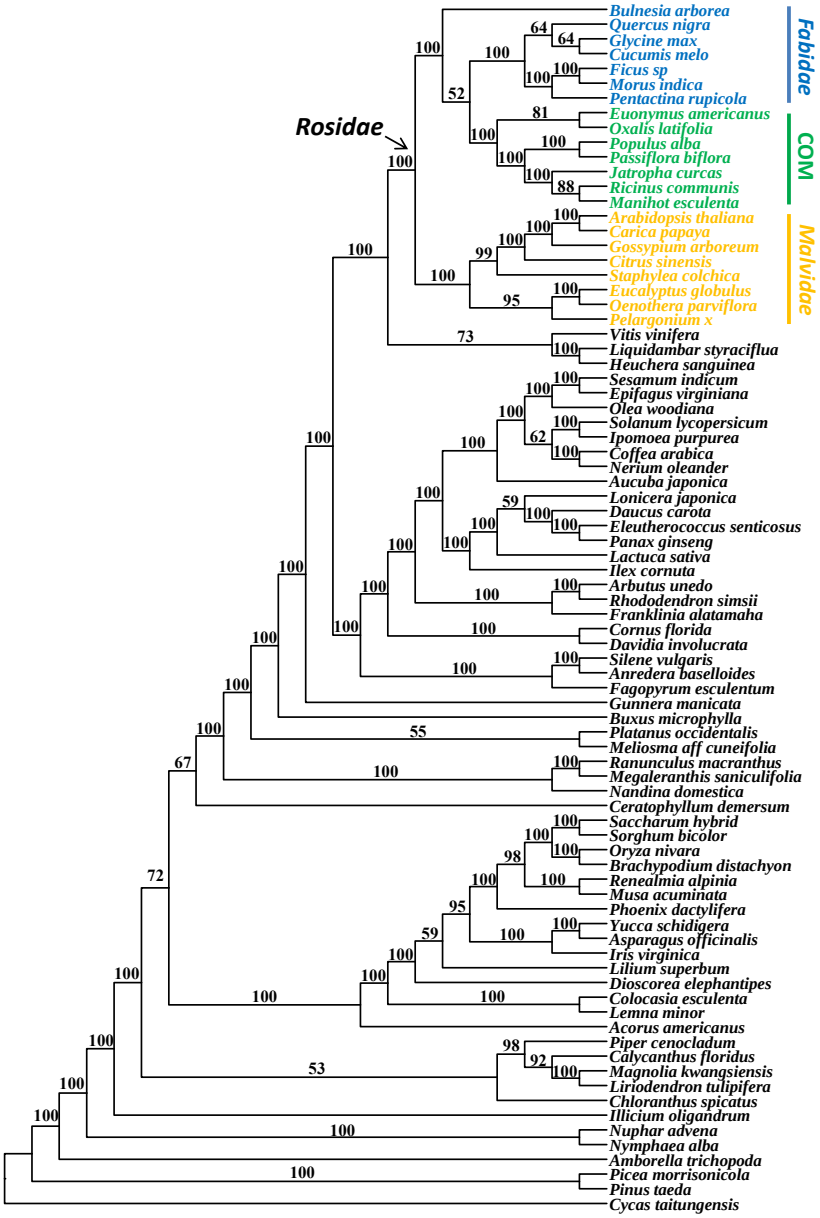


图 3-1 叶绿体 78 基因矩阵的多数一致 ML 树（COM 支与 *Fabidae* 聚类）

Figure 3-1 Maximum likelihood majority-rule consensus bootstrap tree inferred from the nucleotide matrix of 78 chloroplast genes (The COM clade is placed with *Fabidae*)

注：枝上的数字表示 BS 支持率；图中所涉及的科目名称及范围与 APG III（2009）一致；各分支的名称与 Cantino et al.（2007）一致（下同）。

79 类群 4 个线粒体基因的分析结果整体表明 COM 支与 *Malvidae* 类群更近缘的系统发育关系（图 3-2；图 A3）。在线粒体 4 基因的核苷酸矩阵分析中，95% BS 支持 COM 支与 *Malvidae* 大部分类群聚在一起，但不包括 *Stachyurus*（旌节花属，Stachyuraceae，Crossosomatales）和 *Oenothera*（月见草属，Onagraceae，Myrtales）（图 3-2）。此外，*Guaiacum*（愈疮木属，Zygophyllaceae，Zygophyllales）与 *Stachyurus*（Stachyuraceae，Crossosomatales）成姊妹关系（图 3-2）；此系统发育关系与 Qiu et al.（2010）的研究结果一致，而不与前人基于叶绿体基因的研究结论一致。在四个线粒体单基因分析中，仅 *matR* 和 *rps3* 弱支持（BS < 60%）COM + *Malvidae* 的聚类关系，而余下两基因对 COM 支，甚至整个蔷薇类都没有支持（附录 B）。在线粒体氨基酸矩阵分析中，74% BS 支持 COM 支与除 *Stachyurus* 和 *Oenothera* 以外的大部分 *Malvidae* 类群成姊妹关系（图 A3）。在线粒体 RY 编码矩阵分析中，蔷薇类并没有得到很好的支持，COM 支亦无分辨。而在快速进化位点移除分析中，移除 5%的快速进化位点，COM 支与除 *Stachyurus* 和 *Oenothera* 以外的大部分 *Malvidae* 类群的姊妹关系得到 100% BS 支持。但如果移除更多的快速进化位点，整棵树各分支关系的支持率将降低。如，移除快速进化位点的 10%，COM 支的 BS 支持率将降为 23%。

92 类群 5 个核基因的矩阵分析结果与线粒体的分析结果类似（图 3-3）。COM 支与 *Malvidae* 的大部分类群的姊妹关系获得 100% BS 支持，但不包括 *Pelargonium*（天竺葵属，Geraniaceae，Geraniales）、*Oenothera*（Onagraceae，Myrtales）和 *Lagerstroemia*（紫薇属，Lythraceae，Myrtales）。此结论与核基因的 RY 编码和氨基酸序列分析完全一致（图 A4-1，图 A4-2）。在核矩阵的单基因分析中，5 个单拷贝基因都分别支持 COM + *Malvidae* 的关系，但 BS 支持率各自不同（附录 B）。当移除快速进化位点的 5%和 10%后，均 100% BS 支持 COM 支与 *Malvidae* 的大部分类群的姊妹关系获得支持，但不包括 *Pelargonium*、*Oenothera* 和 *Lagerstroemia*。当移除快速进化位点的 20%和 30%后，COM + *Malvidae* 的支持率将分别降为 96%和 94%。但移除更多的进化位点，对蔷薇类以及 COM 支的支持率将大为降低。

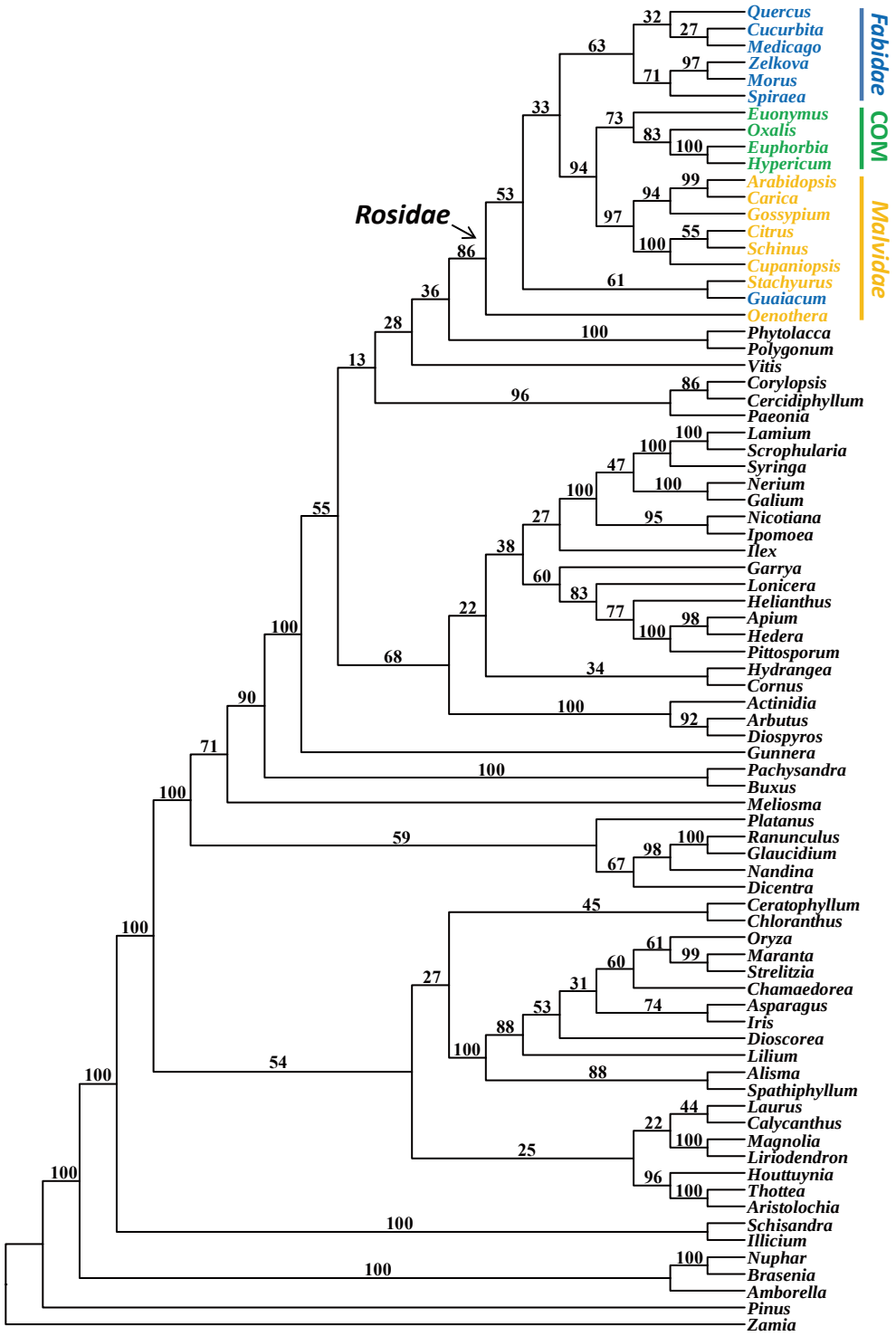


图 3-2 线粒体 4 基因矩阵的多数一致 ML 树（COM 支与 *Malvidae* 聚类）

Figure 3-2 Maximum likelihood majority-rule consensus bootstrap tree inferred from the nucleotide matrix of 4 mitochondrial genes (The COM clade is placed with *Malvidae*)

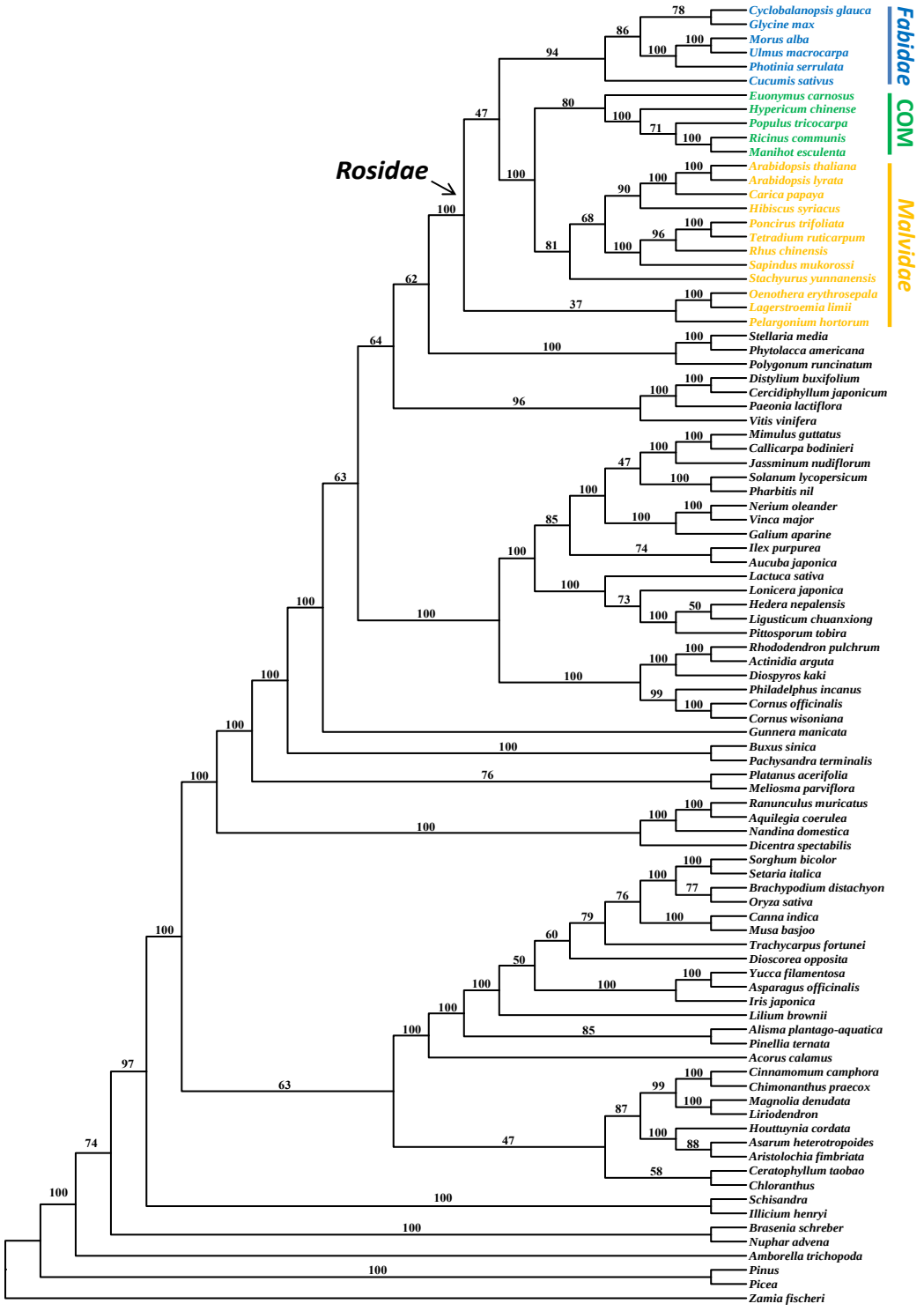


图 3-3 核 5 基因矩阵的多数一致 ML 树（COM 支与 *Malvidae* 聚类）

Figure 3-3 Maximum likelihood majority-rule consensus bootstrap tree inferred from the nucleotide matrix of 5 nuclear genes (These data place the COM clade with *Malvidae*)

3.2 核基因组数据

3.2.1 单拷贝核基因分析结果

尽管 Lee et al.（2011）研究中的大部分单拷贝核基因对于 COM 支系统位置没有信息，但在那些对 COM 支系统关系有信息位点的基因（8,445）中，61–75%单拷贝核基因支持 COM + *Malvidae* 的拓扑关系（表 3-2；图 A5-1），25–39%单拷贝核基因支持 COM +*Fabidae*，但没有任何基因支持 *Fabidae* + *Malvidae* 拓扑关系（表 3-2；图 A5-1）。

表 3-2 基于 Lee et al.（2011）直系同源矩阵的单拷贝核基因分析结果

Table 3-2 Results from the single-copy nuclear gene analysis based on the ortholog alignments from Lee et al. (2011)

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

注：a、表中数字表示在给定 BS 支持率下支持 COM 支对应系统位置的单拷贝核基因数；

b、括号内的数字表示此基因数占所有 8,445 个对 COM 支系统位置有信息位的单拷贝核基因的百分比；

c、“*Fabidae*”代表 COM + *Fabidae* 拓扑关系，“*Malvidae*”代表 COM + *Malvidae* 拓扑关系，“*Fabidae* + *Malvidae*”为 COM 支与 *Fabidae* + *Malvidae* 成姊妹群的拓扑关系（图 2-1）。

整体上，随着 BS 支持率的递增，支持 COM 支三种系统位置的单拷贝核基因的数量都在减少，但是对于支持 COM + *Malvidae* 和 COM +*Fabidae* 拓扑结构的单拷贝核基因相对数量各自均变化不大。详细结果描述如下：1)、随着 BS 支持率的增加，支持前两种拓扑关系（即 COM + *Fabidae* 和 COM + *Malvidae*）的单拷贝基因数目在减少（表 3-2；图 A5-1）；2)、当 BS 值从 10%递增至 90%，支持 COM + *Malvidae* 拓扑关系的单拷贝基因的比例从 61%增加到 65%，而支持 COM + *Fabidae* 拓扑关系的单拷贝基因的比例从 39%减少到 35%（图 A5-1）；3)、当支持率为 100%时，在这些所有的有信息位点的单拷贝基因中，75%的单拷贝核基因支持 COM + *Malvidae* 拓扑关系，而仅有 25%的基因支持 COM + *Fabidae* 的关系（表 3-2；图 A5-1）。

综上所述，在所有对 COM 支系统位置有信息的单拷贝核基因中，绝大部分单拷贝核基因支持 COM + *Malvidae* 系统发育关系，尽管也有小部分一定比例的单拷贝核基因支持 COM + *Fabidae*，而对 *Fabidae* + *Malvidae* 拓扑关系没有任何支持。

3.2.2 多拷贝核基因分析结果

多拷贝核基因的分析结果与单拷贝的分析结果相似：大部分多拷贝核基因对 COM 支的系统位置没有信息，但在那些对 COM 支的系统发育关系有信息的基因中，绝大部分支持 COM + *Malvidae* 的系统发育关系（表 3-3；图 A5-2–A5-4）。在所有三种不同的调和成本（三种进化事件模型）下，71–98%的多拷贝核基因支持 COM + *Malvidae* 的系统关系，而支持 COM + *Fabidae* 和 *Fabidae* + *Malvidae* 的多拷贝基因百分比分别为 1–27%和 0–6%（表 3-3；图 A5-2–A5-4）。

在基因重复事件模型下，91–98%的基因支持 COM + *Malvidae* 的拓扑关系，而仅有 1–6%的基因支持 COM + *Fabidae*（表 3-3；图 A5-2），不足 6%的基因支持 *Fabidae* + *Malvidae*。从图 A5-2 很容易看出，随着支持率从 10%增至 90%，支持 COM + *Fabidae* 基因的百分比降到 1%，而支持 COM + *Malvidae* 的基因的百分比增至 98%。当 BS 支持率为 100%时，支持 COM + *Fabidae* 和 COM + *Malvidae* 的基因的百分比分别为 6%和 94%，而没有任何基因支持 *Fabidae* + *Malvidae* 的拓扑关系。

表 3-3 基于 22 个陆地植物全基因组的多拷贝核基因分析结果

Table 3-3 Results from the multi-copy gene tree analysis based on genome sequences of 22 land plant taxa

进化事件	% BS 支持率	*Fabidae*	*Malvidae*	*Fabidae* + *Malvidae*
基因重复	50	38 (3%)	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%)
基因重复与丢失	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%)
不完全谱系筛选	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%)

注：a、此分析方法是首先对 3,784 多拷贝核基因矩阵进行 ML 分析以获得多拷贝核基因树，然后对这些树在三种进化事件（基因重复、基因重复与丢失和不完全谱系筛选）发生次数最少条件下来统计，并得出调和成本；

b、表中数字表示在 ≥50% BS 支持率和三种不同调和成本下，支持对应 COM 支系统位置的多拷贝核基因数；

c、括号内的数字表示此基因数占所有 3,784 个对 COM 支系统位置有信息位的多拷贝核基因的百分比；

d、“*Fabidae*”代表 COM + *Fabidae* 拓扑关系，“*Malvidae*”代表 COM + *Malvidae* 拓扑关系，“*Fabidae* + *Malvidae*”为 COM 支与 *Fabidae* + *Malvidae* 成姊妹群的拓扑结构（图 2-1）。

在不完全谱系筛选事件模型下，同样地随着 BS 支持率的增加，71–85%的

基因支持 COM + *Malvidae*，13–27%的基因支持 COM + *Fabidae*，而仅 1–2%的基因支持 *Fabidae* + *Malvidae*（表 3-3；图 A5-4）。类似地，随着支持率从 50%递增至 90%，支持 COM + *Fabidae* 的基因的百分比从 27%降至 13%；相比之下，支持 COM + *Malvidae* 的基因的百分比从 71%增至 85%。尽管在支持率为 50%BS 时，不完全谱系筛选模型下支持 COM + *Fabidae* 的基因数目最多，但仅占全部有信息基因的 27%（图 A5-2–A5-4）。

在基因重复和丢失事件模型下，随着 BS 支持率的增加，76–91%的基因支持 COM + *Malvidae*，8–20%的基因支持 COM + *Fabidae*，而支持 *Fabidae* + *Malvidae* 基因的百分比不足 4%（表 3-3；图 A5-3）。类似地，随着支持率从 50%递增至 90%，支持 COM + *Fabidae* 的基因的百分比从 20%减至 8%，而支持 COM + *Malvidae* 的基因的百分比增至 91%。当 BS 支持率为 100%时，支持 COM + *Fabidae* 和 COM + *Malvidae* 的基因的百分比分别为 11%和 89%，而对 *Fabidae* + *Malvidae* 没有任何支持。基因重复和丢失事件模型下，支持 COM + *Malvidae* 的基因数目最多，但随着支持率的变化，其百分比在 76%与 91%之间变动（表 3-3；图 A5-3）。

3.3 总结

通过以上各项数据的分析，我们发现：

1）、在取样近等同的三个矩阵中，叶绿体数据支持 COM + *Fabidae* 的系统发育关系（图 3-1；图 A2），而线粒体和核基因的数据支持 COM + *Malvidae* 的系统关系（图 3-2，图 3-3；图 A3，图 A4）。三个矩阵的各单基因分析结果也与上述结论整体一致（附录 B），尽管部分基因对 COM 支的系统位置或没有信息，或与所在的矩阵支持的系统位置有差别（附录 B）。在 RY 编码和快速进化位点移除分析中，叶绿体数据均支持 COM + *Fabidae* 的拓扑关系，而核矩阵支持 COM + *Malvidae*；尽管线粒体的 RY 矩阵对 COM 支的系统位置没有支持，但在移除快速进化位点分析中，移除 5%的快速位点后，线粒体矩阵也是 100% BS 支持 COM + *Malvidae* 的关系。线粒体与核的氨基酸矩阵分别以 74% BS 和 100% BS 支持 COM + *Malvidae*，而叶绿体的氨基酸矩阵仅 47% BS 支持 COM + *Fabidae*。

2）、单拷贝和多拷贝核基因数据的分析结果基本一致，均支持支持 COM + *Malvidae*（表 3-2，表 3-3）。虽然大部分基因对 COM 支的系统位置没有信息，

但是对其位置有信息的基因分析中，61–75%单拷贝核基因支持 COM + *Malvidae* 的拓扑关系，而 71–98%的多拷贝核基因在所有的进化事件模型下都一致支持 COM + *Malvidae*；而支持 COM + *Fabidae* 的基因虽占少数但比列显著，此类单拷贝与多拷贝的基因比列区间分别为 25–39%和 1–27%；对 *Fabidae* + *Malvidae* 关系几乎没有支持，不足 6%（图 A5）。

3）、在多拷贝核基因中，三种进化事件模型都显著支持 COM + *Malvidae* 的系统位置（图 A5-2–A5-4）。其中，基因重复事件模型对 COM + *Malvidae* 的支持率最高。尽管在基因重复事件模型中，对支持 COM + *Fabidae* 基因的百分比不足 6%，但是在另外两个模型中，支持 COM + *Fabidae* 基因的百分比有所增加。这些数据表明 COM + *Fabidae* 的系统发育信号不能忽略（图 A5-2–A5-4）。

第四章 COM 支系统发育关系冲突原因探讨

4.1 冲突原因非取样和系统误差

尽管前人在其研究中单独或联合地应用了不同基因各种组合的数据，也产生了许多新的数据，采用了不同的取样策略、不同的分析方法，但是 COM 支的系统位置仍然不确定（表 1-1）。因此很难判断 COM 支系统位置的冲突是来自系统误差还是生物过程。针对此，本研究设计了代表植物三个基因组的、取样近等同的矩阵（82 类群 78 基因的叶绿体矩阵，79 类群 4 基因的线粒体矩阵和 92 类群 5 基因的核矩阵）以降低取样偏差的影响；用氨基酸和 RY 编码矩阵可以有效地降低核苷酸序列中的置换饱和、进化速率和碱基组成成份的异质性，以提高树的可信度（Hashimoto et al. 1995; Phillips and Penny 2003; Harrison et al. 2004; Delsuc et al. 2005; Gibson et al. 2005）；还采用了快速进化位点移除法以减少长枝吸引和模型错配对所得拓扑结构的影响（Goremykin et al. 2010; Philippe et al. 2005）。尽管应用了上述多种性状编码和“噪音”排除的方法来消除系统误差的影响，但我们的最终研究结果仍重现了两种相互冲突的 COM 支系统位置（图 A1），即叶绿体基因支持 COM + *Fabidae*（图 3-1），而线粒体与核基因支持 COM + *Malvidae*（图 3-2，图 3-3）。此结论进一步加强了对叶绿体基因与线粒体和核基因的结论冲突的观察（图 A1；表 1-1），而且，在对三个基因组的矩阵分别进行快速变异位点的移除时，我们发现随着快速变异位点的移除，要么使拓扑结构的支持率降低，要么使得整个棵树的结构倒塌，并没有新的拓扑结构出现。这说明，矩阵中不存在大量的导致非同源相似的快速进化位点，相反，矩阵中的大部分位点对 COM 支系统位置是有信息的。

针对上述取样偏差和系统误差的检测和分析，我们认为所得到三个基因组的系统发育关系结果是可靠的。尽管上述一系列性状编码和“噪音”排除的方法没有找到系统误差，但也不能绝对排除它的存在。事实上，在连锁的叶绿体单基因分析中，COM 支系统位置出现一些弱支持的变化，这反映出叶绿体数据中有少量误差。另外，在本研究的线粒体与核基因数据分析结果中，我们观察到 *Stachyurus*，*Pelargonium*，*Lagerstroemia* 和 *Oenothera* 出现在孤立且弱支持的

系统位置上，导致 *Malvidae* 不总为单系，但这些类群在本研究中系统位置，分别与它们在 Qiu et al.（2010）和 Zhang et al.（2012）研究中对应的系统位置是一致的。而且，在目前叶绿体、线粒体以及核基因的单独或联合的研究中，这四个属分别所在的桃金娘目（Myrtales）、Geraniales（牻牛儿苗目）以及流苏子目（Crossosomatales）本身的系统位置在蔷薇类内并没有最终固定（如，Morton 2011; Qiu et al. 2010; Soltis et al. 2011; Zhang et al. 2012; Zhu et al. 2007）。总之，三个矩阵的分析结果仍然重现了两种 COM 支拓扑关系，这种与前人研究所揭示的冲突是一致的，说明关于 COM 支系统位置在叶绿体与线粒体和核基因间的冲突是客观存在的。这暗示在叶绿体、线粒体和核三个基因组的基因座间潜藏着以生物过程为基础的冲突。

4.2 冲突信号与核基因组数据分析

关于 COM 支的系统位置，如果说叶绿体、线粒体以及核基因数据间的冲突是由生物进化事件导致的话，那么这种冲突也会在双亲遗传、非连锁且独立的核基因组内观察到。

事实上，在 Lee et al.（2011）单拷贝直系同源核基因矩阵的分析中，我们发现 61–75%的单拷贝基因支持 COM + *Malvidae* 的系统发育关系，而 25–39%的单拷贝基因支持 COM + *Fabidae* 的系统发育关系（表 3-2）。基于三种拓扑结构（图 2-1），我们对 8,445 个单拷贝核基因矩阵分析发现，随着支持率从 10%到 100%，支持 COM + *Fabidae* 的拓扑关系的基因比例在 25–39%范围内变化，而支持 COM + *Malvidae* 的系统发育关系的基因比例在 61–75%范围内变化（图 A5-1）。二者的相对比例并没有明显的差异。因此，我们认为这两种相互冲突的系统发育信号并存的现象可能与不完全谱系筛选或与蔷薇类内古杂交事件有关，而核基因组内 COM + *Fabidae* 的系统发育信号可能是其中一个亲本遗传信息的残留（详见 4.3 节）。

在多拷贝核基因矩阵，我们也同样发现了 COM 支冲突的系统位置。在这些多拷贝核基因中，至少 71%的多拷贝核基因支持 COM + *Malvidae* 的系统发育关系，而仅有 1–27%的多拷贝基因支持 COM + *Fabidae*，对 *Fabidae* + *Malvidae* 关系几乎没有支持（表 3-3；图 A5-2–A5-4）。这种核基因组内 COM + *Malvidae* 的系统发育关系占主导的结论，与 Burleigh et al.（2011）和 Górecki et al.（2012）

的基因树简约分析法结论一致。而且在核基因组内，在所有的进化事件模型下，都支持 COM + *Malvidae* 的系统发育关系，也说明此 COM 支系统关系是可靠的。

总之，单拷贝与多拷贝核基因组数据分析中，多数对 COM 支有信息位点的核基因一致支持 COM + *Malvidae* 的系统发育关系，尤其是只考虑对拓扑结构支持率最高的基因时，这种一致性更显著。而且，在单拷贝和多拷贝核基因的分析结果一致显示，不管 BS 支持率在哪个区间变化，而支持 COM 支三种可能拓扑结构基因的总体比例是相对稳定的（图 A5）。另外，在核基因组分析中，还表明基因重复事件也可能是造成核基因组中存在两种不同系统发育信号的原因（旁系同源）。核基因组通常被认为是由包含不同进化历史的核苷酸序列组成的“马赛克”，而且核基因组的基因重复和/或丢失事件使得旁系同源基因的存在在也很普遍。故，在核基因中检测到不同的系统发育信号也是很有可能的。当然，鉴于取样有限以及基因组注释或基因家族界定的错误的原因，也可能导致潜在的基因序列的遗漏和基因树拓扑结构估计错误。所以，精确估计出基因重复、基因重复和丢失以及谱系筛选事件发生的次数几乎是不可能的（Hahn 2007; Rasmussen and Kellis 2011）。

4.3 生物过程导致 COM 支系统位置冲突

通过上一节的分析，我们知道 COM 支系统位置冲突如果不是由取样和系统误差的引起的话，那就是能够引起叶绿体、线粒体和核三个基因组的基因座间产生差异的生物过程造成的。基于 COM 支系统位置，叶绿体与线粒体数据的冲突说明二者在此类群的遗传上不连锁，分别具有不同遗传历史。一般来说，两种进化过程最有可能造成线粒体和叶绿体不同的遗传历史，从而影响到 COM 支的系统位置：不完全谱系筛选和伴随叶绿体渐渗的古杂交事件。那么在蔷薇类快速分化的早期，在 4–5 Mya 的相对狭窄时间范围内，是古杂交还是不完全谱系筛选事件造成 COM 支的系统发育位置在三个基因组间的不一致呢？

4.3.1 不完全谱系筛选假说

一般来说，祖先物种的全部基因都会传递给后续分化的所有后代。但随着时间的推移，祖先的基因由于自然选择或遗传漂变等因素未能完全地传递给其

所有后代，部分基因在某些谱系中丢失了，造成不同谱系间的基因不一致，即不完全谱系筛选。由此，导致基于这些基因的推导的系统发育关系产生冲突，而且不同基因造成冲突的格局也不一样。类群的快速辐射分化能够加速因不完全谱系筛选造成的基因或者基因组间的进化历史不一致（Enard and Paabo 2004; Pollard et al. 2006; Whitfield and Lockhart 2007; Oliver 2013），而且当有效种群相当大时，不完全谱系筛选更容易发生（Pamilo and Nei 1988）。尽管不完全谱系筛选主要会影响近期的辐射分化事件，但早期的快速辐射分化也有可能诱发不完全谱系筛选，从而影响到系统发育关系更高阶元的、更深次的系统发育关系的解决（Whitfield and Lockhart 2007; Murphy et al. 2007; Degnan and Rosenberg 2009）。比如，在哺乳动物中深层次的系统发育关系冲突就与此有关（McCormack et al. 2012; Song et al. 2012）。蔷薇类中的两大亚支 *Fabidae* 和 *Malvidae* 估计的最早分化时间分别为 108–91 Mya 与 107–83 Mya，而且在不到 4–5 Mya 短暂的时间内，蔷薇类迅速完成了物种分化（Wang et al. 2009），那么蔷薇类内 COM 支系统位置的冲突就有可能是蔷薇类在快速辐射分化早期的不完全谱系筛选事件导致的。

目前已有相当多的研究表明，区分不完全谱系筛选和杂交事件很具挑战性（如，Sang and Zhong 2000; Holder et al. 2001; Buckley et al. 2006; Holland et al. 2008; Joly et al. 2009）。再者，本研究中核基因组数据的取样稀疏、不完整以及久远的蔷薇类分化时间（Wang et al. 2009; Bell et al. 2010）等这些因素，使得区别上述二者更加困难。不完全谱系筛选在物种或居群水平上通常发生的概率更大一些（Comes and Abbott 2001; Xiang et al. 2005; Maddison and Knowles 2006; Pollard et al. 2006; Zou et al. 2008; Yu et al. 2011）。而且，我们若把 COM 支与 *Fabidae* 和 *Malvidae* 之间的系统发育关系简化为三个分类群有根的系统发育关系问题来看的话（图 2-1），则依据 Huson et al.（2005）的模型理论，如果发生了不完全谱系筛选事件，那么在仅有的三种拓扑结构中，除物种树的拓扑结构之外，支持另外两种拓扑结构的核基因数应该是等量的。然而，我们的分析结果却与此恰恰相反，大部分基因支持 COM + *Malvidae*，仅少数基因支持 COM + *Fabidae*，而对 *Fabidae* + *Malvidae* 几乎没有支持（表 3-2，表 3-3）。从支持这三种可能拓扑关系的基因比例来看，此格局并不支持用不完全谱系筛选的假说来解释不同基因间的系统发育关系冲突。不过，鉴于 *Malvidae* 和 *Fabidae* 在我们的某些分析中并没有完全聚为一支（图 3-1–图 3-3），也有可能 Huson et al.（2005）

的“三个类群有根的系统发育关系”理论在这里并不完全适用。而大于三个类群的不完全谱系筛选理论远比此复杂（如，Rosenberg and Nordberg 2002; Degnan and Rosenberg 2006）。且近期 Reid 等基于模型匹配的分析表明多类群的不完全谱系筛选原理并不适用于大多数系统发育数据的分析（Reid et al. 2013）。

4.3.2 古杂交假说

许多植物类群在其繁衍史中都经历过多次杂交或者叶绿体渐渗事件（Okuyama et al. 2005），而蔷薇类植物在 20 多年前就有 100 多例叶绿体 DNA 渐渗事件的报道（Rieseberg and Soltis 1991; Rieseberg et al. 1996a）。物种间叶绿体基因或基因组渐渗现象在植物各分类阶元水平上均有可能发生，尤其是在种、属级水平上，物种杂交并伴随叶绿体渐渗事件造成叶绿体基因与其他基因间所得系统发育关系不一致的案例很普遍（Soltis and Kuzoff 1995; Wendel et al. 1995; Rieseberg et al. 1995, 1996a, 1996b; Raamsdonk et al. 1997; Peters et al. 2007）。

伴随叶绿体渐渗现象的古杂交事件同样也会造成独立基因座间的系统发育关系冲突。基于 COM 支的系统位置，从目前分析来看，叶绿体基因支持 COM + *Fabidae*，而线粒体支持 COM + *Malvidae*，说明该类群这两个亚细胞器单元（subcellular compartments）在遗传上并不关联，而且具有各自不同遗传进化历史，即叶绿体基因组源自 *Fabidae* 的祖先谱系，而线粒体是来自 *Malvidae* 的祖先谱系。但这样的结论有些意外，因为，众所周知，在被子植物中叶绿体和线粒体基因组都是典型的母系遗传（Birky 1995, 2001; Corriveau and Coleman 1988; Mogensen 1996）。虽然，我们目前很难评估蔷薇类祖先类群的叶绿体与线粒体基因组的遗传模式，但目前已有不少叶绿体与线粒体分别遗传自两个不同亲本的研究报道（如，Fauré et al. 1994; Havey et al. 1998; Testolin and Cipriani 1997; Yang et al. 2000），而且还有一些蔷薇类内叶绿体基因组父系遗传的报道，如来自 COM 支类群（时钟花 *Turnera ulmifolia*; Shore et al. 1994; Shore and Triassi 1998）和 *Fabidae* 类群（紫苜蓿 *Medicago sativa*; Schumann and Hancock 1989; Masoud et al. 1990; 石碳酸灌木属 *Larrea*; Yang et al. 2000）。另外，还有实验证据表明，植物在其杂交后代中会体现出显著的叶绿体父系遗传而线粒体严格母系遗传的现象（Schumann and Hancock 1989; Masoud et al. 1990; Shore et al. 1994; Xu 2005）。因此，通过蔷薇类内的古杂交事件导致叶绿体与线粒体基因组具有不同的进化历

史假说是非常有可能的。

在这个古杂交假说中，*Fabidae* 的祖先谱系（或其直接祖先）充当了父本与母本 *Malvidae* 的祖先谱系进行了杂交。在此杂交过程中伴随着叶绿体父系遗传现象，即其父本 *Fabidae* 将其叶绿体基因组传递给 COM 支的祖先（F_1）。此生物过程使得 COM 支（F_1）的叶绿体和线粒体具有不同的进化历史，导致了 COM 支（F_1）内来自两基因组的基因树相互冲突，同样也导致了核基因组内因双亲等位基因的存在而引起的冲突（图 4-1；图 A-1）。然而上述类群间单纯的杂交或自交事件并不能解释核基因组内高比例的核基因支持 COM + *Malvidae* 系统关系（表 3-2，表 3-3），这是因为早期杂交后的子代又与其原来的母本 *Malvidae* 发生了回交（backcrosse）事件，最终致使 COM 支内 COM + *Malvidae* 的遗传信息得到了累积，而支持 COM + *Fabidae* 得到了稀释或消减。所以，这样最终造成了 COM 类群中的叶绿体与线粒体基因的冲突，以及核基因组内不等比例且代表着两个亲本遗传信息的系统发育信号的存在（图 4-1）。

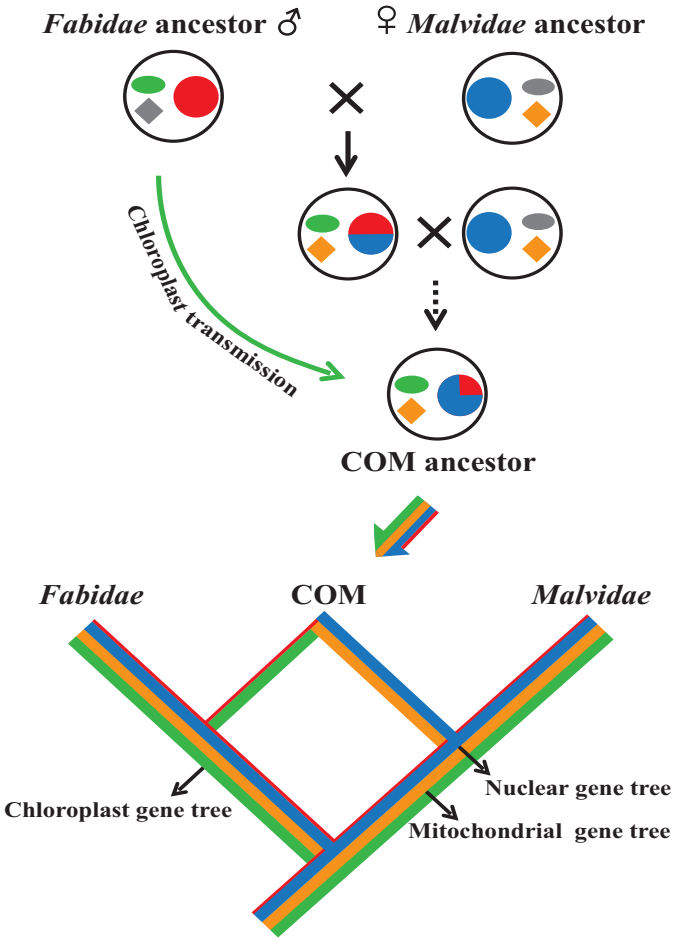


图4-1 蔷薇类内辐射分化早期*Fabidae*与*Mavidae*的祖先谱系杂交产生COM支的假说图示

Figure 4-1 Hypothetical reticulation scenario for the origin of the COM clade from the ancestral *Fabidae* and *Malvidae* lineages

注：图中的大圆代表植物谱系；小圆代表不同谱系的核DNA（红色代表*Fabidae*祖先谱系的核DNA，蓝色代表*Mavidae*祖先谱系的核DNA）；椭圆代表叶绿体细胞器（绿色代表来自*Fabidae*祖先谱系，灰色代表来自杂交前的*Mavidae*祖先谱系）；菱形代表线粒体细胞器（灰色代表来自*Fabidae*祖先谱系，橙色代表来自*Mavidae*祖先谱系）；绿色箭头表示在杂交过程叶绿体父系遗传（叶绿体渐渗）；虚线箭头代表杂交子代与其母本数代的回交。在此杂交过程中，线粒体为母系遗传，故COM支的祖先（F_1）的线粒体直接来自其母本*Mavidae*的祖先；而其叶绿体为父系遗传，故COM支的祖先（F_1）的叶绿体来自其父本*Fabidae*的祖先。此后，F_1代与其母本*Mavidae*多次回交。因此，这最终导致了COM支的后代中的叶绿体来自*Fabidae*，线粒体来自*Mavidae*，而大部分的核基因来自*Mavidae*，而小部分（约25%）来自*Fabidae*。这样复杂的网状进化历程最终导致了图底部三个基因组间冲突的系统发育关系。

4.4 结论与展望

本研究利用代表植物基因组的三个矩阵进行分析说明造成COM支叶绿体与线粒体和核基因数据冲突的根本原因是复杂的生物过程，即在蔷薇类起源和分化的早期，*Fabidae*和*Malvidae*的祖先谱系可能发生了杂交并伴随叶绿体基因组渐渗的进化事件。此生物事件最终导致了COM支类群在叶绿体和线粒体内蕴含了不一致的遗传信息，以及核基因组内不等比例地保存两亲本不同的遗传信息。

诸多的植物系统学研究表明利用基因组数据能够揭示一些传统的若干基因研究不能够捕获的、深层次的被子植物系统发育信息（Finet et al. 2010; Moore et al. 2010, 2011; Burleigh et al. 2011; Lee et al. 2011）。本研究正是基于多个不连锁的基因组数据分析再次揭示了被子植物系统发育关系中不明晰、甚至冲突的节点，并挖掘了潜藏在冲突背后复杂进化历史。大部分单拷贝及多拷贝核基因和线粒体基因都持COM支与*Malvidae*的近缘关系，与形态性状的进化模式也相吻合（Endress and Matthews 2006; Endress et al. 2013），但与叶绿体数据高度支持的结论相悖（图A-1; Jansen et al. 2007; Moore et al. 2010, 2011; Ruhfel et al. 2014）。由此说明，与其仅依靠叶绿体基因组的数据来寻求唯一的系统发育关系，强行把COM支的系统位置定在一棵二歧分支的物种树上，还不如结合其他源数据来尝试探讨COM支不同的起源，这样才更具信息量、更有进化意义。尽管目前的类群和基因的取样，还不足以对COM支来源和所涉及的进化事件下定论，但本

研究着重体现了利用基因组数据揭示被子植物深层次的系统发育关系冲突的重要意义，并为本领域将来开展的其他研究提供了参考。

联合大量基因的系统发育分析方法可能会掩盖导致系统发育关系冲突的信号，人们长期以来一直受此困扰。因为这种信号往往可能是生物过程导致的（Soltis et al. 2004; Pollard et al. 2006; Beiko et al. 2008; Maureira-Butler et al. 2008; Salichos and Rokas 2013）。尽管不完全谱系筛选、杂交等网状进化过程被认为是在植物种和居群水平上最要的进化动力，但是本研究表明它们的潜在影响力还涉及到被子植物的更深层次。类似地，Oliver（2013）利用谱系建模的方法也揭示了微进化在生命之树上也起着相似的作用，但未得人们到充分关注。蔷薇类是被子植物重要的分支之一，其进化背景复杂，它可能在不到4–5 Mya短暂的时间内产生了快速辐射分化（Wang et al. 2009）。本研究首次在被子植物系统发育关系的深层次上提出了古老的进化事件造成系统发育关系冲突的假说，可为被子植物其他类群经历不完全谱系筛选、杂交等生物过程造成的系统发育关系冲突的揭示有一定的参考作用。

通过本文前面各章节分析、总结可见，揭示生物进化的动态过程可能是澄清生命之树疑难节点所面临的首要课题，而且有些类群所涉及的进化事件可能不是单一的。杂交/渐渗、不完全谱系筛选往往是造成物种系统发育关系冲突的两大主要因素。本研究不仅对解决生命之树上其他疑难节点的系统发育关系方面具有一定的借鉴意义；而且还展示了系统发育基因组学方法在检测、识别以及解决生物进化事件导致基因树的冲突方面是行之有效的。与此同时，在利用海量的分子数据解析类群系统发育关系时，我们还应当更谨慎一些。鉴于叶绿体、线粒体与核三个植物细胞器基因组不同的遗传模式和进化背景，仅仅依赖于少数基因或个别基因组来重建系统发育关系有时是不可靠的（如，COM 支），即使 100%支持率的单基因树也并不能保证真实、可靠的系统发育关系，更重要的是还可能会遗漏潜藏的生物进化过程。

毫无疑问，随着测序技术蓬勃发展和基因组时代的到来，不同来源的数据海量增加，利用这些数据得出的系统发育关系的冲突现象会越来越频繁，且不可避免。存在系统发育关系冲突，就暗示着这些数据可能没有真实地反应类群的系统发育关系。理论上讲，来自不同基因、基因座的数据得出的系统发育关系越相似，越能反映真实的物种系统发育关系，越接近物种树（Qiu et al. 2010）。因此，选用不同基因组的、进化速率均衡的数据，以及采用不同的方法来重建

系统发育关系很重要。而且，结合不同基因组的数据、非分子数据（如形态、化石等）来相互验证也很必要。唯有如此，才能进一步揭露数据可能存在的错误、或非同源相似、或潜藏的生物进化过程，而最终获得真实、可信的系统发育关系。

参考文献

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 I (1998) An ordinal classification for the families of flowering plants. *Ann. Mo. Bot. Gard.* **85,** 531–553.

APG II (2003) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. *Bot. J. Linn. Soc.* **141,** 399–436.

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.

Baldwin BG (1992) Phylogenetic utility of the internal transcribed spacers of nuclear ribosomal DNA in plants: An example from the Compositae. *Mol. Phylogenet. Evol.* **1,** 3–16.

Bansal MS, Burleigh JG, Eulenstein O, Wehe A (2007) Heuristics for the gene-duplication problem: A Θ(n) speed-up for the local search. In: *Research in computational molecular biology*. Heidelberg, Springer, pp. 238–252.

Beiko RG, Doolittle WF, Charlebois RL (2008) The impact of reticulate evolution on genome phylogeny. *Syst. Biol.* **57,** 844–856.

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.

Brinkmann H, Philippe H (1999) Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. *Mol. Biol. Evol.* **16,** 817–825.

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, Mathews S (2004) Phylogenetic signal in nucleotide data from seed plants: Implications for resolving the seed plant tree of life. *Am. J. Bot.* **91,** 1599–1613.

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, Kron KA, Rettig JH, Conti E, Palmer JD, Manhart JR, Sytsma KJ, Michaels HJ, Kress WJ, Karol KG, Clark WD, Hedren M, Gaut BS, Jansen RK, Kim KJ, Wimpee CF, Smith JF, Furnier GR, Strauss SH, Xiang QY, Plunkett GM, Soltis PS, Swensen SM, Williams SE, Gadek PA, Quinn CJ, Eguiarte LE, Golenberg EJr, Learn GH, Graham SW, Barrett SCH, Dayanandan S, Albert VA (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, Givnish TJ, Sytsma KJ, Pires JC (2000) Higher-level systematics of the monocotyledons: An assessment of current knowledge and a new classification. In: *Monocots: systematics and evolution* (eds. **Wilson KL, Morrison DA**). Australia, Victoria, Collingwood, CSIRO Publishing, pp. 3–16.

Chase, MW, Fay MF, Devey DS, Maurin O, Rønsted N, Davies TJ, Pillon Y, Petersen G, Serberg O, Tamura MN, Asmussen CB, Hilu K, Borsch T, Davis JI, Stevenson DW, Pires JC, Givnish TJ, Systsma KJ, Mcpherson MA, Graham SW, Rai HS (2006). Multigene analyses of monocot relationships: A summary. *Aliso* **22,** 63–75.

Chen F, Mackey AJ, Stoeckert, CJJr, 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.

Darwin C (1859) On the origin of species by means of natural selection. London: J. Murray.

Davis CC, Wurdack KJ (2004) Host-to-parasite gene transfer in flowering plants: Phylogenetic evidence from Malpighiales. *Science* **305,** 676–678.

Degnan JH, Rosenberg NA (2006) Discordance of species trees with their most likely gene trees. *PLoS Genet.* **2,** 762–768.

Degnan JH, Rosenberg, NA (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* **24,** 332–340.

deQueiroz A, Donoghue MJ, Kim J (1995) Separate versus combined analysis of phylogenetic evidence. *Annu. Rev. Ecol. Evol. Syst.* **26,** 657–681.

Delsuc F, Phillips MJ, Penny D (2003) Comment on "Hexapod origins: Monophyletic or paraphyletic?". *Science* **301,** 1482.

Delsuc F, Brinkmann FH, Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.* **6,** 361–375.

Dial KP, Marzluff JM (1989) Nonrandom diversification with in taxonomic assemblages. *Syst. Biol.* **38,** 26–37.

Doyle JJ (1992) Gene trees and species trees – molecular systematics as one-character taxonomy. *Syst. Bot.* **17,** 144–163.

Drew BT, Ruhfel BR, Smith SA, Moore MJ, Briggs BG, Gitzendanner MA, Soltis PS, Soltis DE (2014) Another look at the poot of the angiosperms reveals a familiar tale. *Syst. Bot.* doi:10.1093/sysbio/syt108.

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, Sørensen MV, Haddock SHD, Schmidt-Rhaesa A, Okusu A, Kristensen RM, Wheeler WC, Martindale MQ, Giribet G (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* **452,** 745–749.

Enard W, Paabo S (2004) Comparative primate genomics. *Annu. Rev. Genomics Hum. Genet.* **5,** 351–378.

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.

Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27,** 401–410.

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.

Frohlich MW, Chase MW (2007) After a dozen years of progress the origin of angiosperms is still a great mystery. *Nature* **450,** 1184–1189.

Galtier N, Daubin V (2008) Dealing with incongruence in phylogenomic analyses. *Philos. Trans. Soc. Lond. B. Bio. Sci.* **363,** 4023–4029.

Gibson TC, Kubisch HM, Brenner CA (2005) Mitochondrial DNA deletions in rhesus macaque oocytes and embryos. *Mol. Hum. Reprod.* **11,** 785–789.

Givnish TJ, Pires JC, Graham SW, McPherson MA, Prince LM, Patterson, TB Rai, HS Roalson ER, Evans TM, Hahn WJ, Millam KC, Meerow AW, Molvray M, Kores P, O'Brien HE, Kress WJ, Hall J, Sytsma KJ (2006) Phylogeny of the monocotyledons based on the highly informative plastid gene *ndhF*: Evidence for widespread concerted convergence. In: *Monocots: Comparative biology and evolution (Excluding Poales)* (eds. **Columbus JT, Friar EA, Porter JM, Prince LM, Simpson MG**). California, Claremont, Rancho Santa Ana Botanic Garden, pp. 28–51.

Givnish TJ, Ames M, McNeal JR, dePamphilis CW, Graham SW, Pires JC, Stevenson DW, Zomlefer WB, Briggs BG, Duvall MR, Moore MJ, Heaney JM, Soltis DE, Soltis PS, Thiele K, Leebens-Mack JH (2010) Assembling the tree of the monocotyledons: Plastome sequence phylogeny and evolution of Poales. *Ann. Mo. Bot. Gard.* **97,** 584–616.

Goloboff PA, Catalano SA, Mirande JM, Szumik CA, Arias JS, Källersjö M, Farris JS (2009) Phylogenetic analysis of 73,060 taxa corroborates major eukaryotic groups. *Cladistics* **25,** 211–230.

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, Heidelberg, Springer, pp. 102–114.

Goremykin VV, Hirsch-Ernst KI, Wölfl S, Hellwig FH (2003) Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm. *Mol. Biol. Evol.* **20,** 1499–1505.

Goremykin VV, Hirsch-Ernst KI, Wölfl S, Hellwig FH (2004) The chloroplast genome of *Nymphaea alba*: Whole-genome analyses and the problem of identifying the most basal angiosperm. *Mol. Biol. Evol.* **21,** 1445–1454.

Goremykin VV, Viola R, Hellwig FH (2009) Removal of noisy characters from chloroplast genome-scale data suggests revision of phylogenetic placements of *Amborella* and *Ceratophyllum*. *J. Mol. Evol.* **68,** 197–204.

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

Goremykin VV, Nikiforova SV, Biggs PJ, Zhong B, DeLange P, Martin W, Woetzel S, Atherton RA, McLenachan T, Lockhart PJ (2013) The evolutionary root of flowering plants. *Syst. Biol.* **62,** 50–61.

Graham SW, Zgurski JM, McPherson MA, Cherniawsky DM, Saarela JM, Horne ESC, Smith SY, Wong WA, O'Brien HE, Pires JC, Olmstead RG, Chase MW, Rai HS (2006) Robust inference of monocot deep phylogeny using an expanded multigene plastid data set. *Aliso* **22,** 3–20.

Gusfield D, Bansal V (2005) A fundamental decomposition theory for phylogenetic networks and incompatible characters. In: *Research in computational biology: Lecture notes in computer science* **(**eds. **Miyano S, Mesirov J, Kasif S, Istrail S, Pevzner P, Waterman M)**. Berlin, Springer-Verlag, pp. 217–232.

Hahn MW (2007) Bias in phylogenetic tree reconciliation methods: Implications for vertebrate genome evolution. *Genome Biol.* **8,** R141.

Hamilton JP, Buell CR (2012) Advances in plant genome sequencing. *Plant J.* **70,** 177–190.

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.

Hasegawa M, Hashimoto T (1993) Ribosomal RNA tree misleading? *Nature* **361,** 23.

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.

Hillis DM (1996) Inferring complex phylogenies. *Nature* **383,** 130–131.

Hilu KW, Borsch T, Müller K, Soltis DE, Soltis PS, Savolainen V, Chase MW, Powell MP, Alice LA, Evans R, Sauquet H, Neinhuis C, Slotta TAB, Rohwer JG, Campbell CS, Chatrou LW (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 B, Moulton V (2003) Consensus networks: A method for visualising incompatibilities in collections of trees. In: *Algorithms in bioinformatics* (eds. **Benson G, Page R**). Berlin, Springer-Verlag, pp. 165–176.

Holland BR, Huber KT, Moulton V, Lockhart PJ (2004) Using consensus networks to visualize contradictory evidence for species phylogeny. *Mol. Biol. Evol.* **21,** 1459–1461.

Holland BR, Jermiin LS, Moulton V (2006) Improved consensus network techniques for genome-scale phylogeny. *Mol. Biol. Evol.* **23,** 848–855.

Holland BR, Benthin S, Lockhart PJ, MoultonV, Huber KT (2008) Using supernetworks to distinguish hybridization from incomplete lineage sorting. *BMC Evol. Biol.* **8 20**2.

Hong DY, Chen ZD, Qiu YL, Donoghue MJ (2008) Tracing patterns of evolution through the tree of life: Introduction. *J. Syst. Evol.* **46,** 237–238.

Huang H, Knowles LL (2009) What is the danger of the anomaly zone for empirical phylogenetics? *Syst. Biol.* **58,** 527–536.

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, pp. 233–249.

Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23,** 254–267.

Hudson RR (1983) Testing the constant-rate neutral allele model with protein sequence data. *Evolution* **37,** 203–217.

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.

Jeffroy O, Brinkmann H, Delsuc F, Philippe H (2006) Phylogenomics: The beginning of incongruence? *Trends Genet.* **22,** 225–231.

Jerrold I, Davis DW, Stevenson GP, Ole S, Lisa M, Campbell JV, Freudenstein DH, Goldman CR, Hardy FA, Michelangeli MP, Simmons CD, Specht FVS, Maria G (2004) 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.

Jian SG, Soltis PS, Gitzendanner MA, Moore MJ, Li RQ, Hendry TA, Qiu YL, Dhingra A, Bell CD, Soltis DE (2008) Resolving an ancient, rapid radiation in Saxifragales. *Syst. Biol.,* **57,** 38–57.

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.

Källersjö M, Farris JS, Chase MW, Bremer B, Fay MF, Humphries CJ, Petersen G, Seberg O, Bremer K (1998) Simultaneous parsimony jackknife analysis of 2538 *rbcL* DNA sequences reveals support for major clades of green plants, land plants, seed plants and flowering plants. *Plant Syst. Evol.* **213,** 259–287.

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.

Kubatko LS, Degnan JH (2007) Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst. Biol.* **56,** 17–24.

Lee EK, Cibrian-Jaramillo A, Kolokotronis SO, Katari MS, Stamatakis A, Ott M, Chiu JC, Little DP, Stevenson DW, McCombie WR, Martienssen RA, Coruzzi G, DeSalle R (2011) A functional phylogenomic view of the seed plants. *PLoS Genet.* **7,** e1002411.

Linder C R, Rieseberg LH (2004) Reconstructing patterns of reticulate evolution in plants. *Am. J. Bot.* **9,** 1700–1708.

Lopez P, Casane D, Philippe H (2002) Heterotachy, an important process of protein evolution. *Mol. Biol. Evol.* **19,** 1–7.

Loomis W F, Smith D W (1990) Molecular phylogeny of *Dictyostelium discoideum* by protein sequence comparison. *Proc. Natl. Acad. Sci. USA* **87,** 9093–9097.

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.

Mathews S, Donoghue MJ (1999) The root of angiosperm phylogeny inferred from duplicate phytochrome genes. *Science* **286,** 947–950.

Matsen FA, Steel M (2007) Phylogenetic mixtures on a single tree can mimic a tree of another topology. *Syst. Biol.* **56,** 767–775.

Maureira-Butler IJ, Pfeil BE, Muangprom A, Osborn TC, Doyle JJ (2008) The reticulate history of *Medicago* (Fabaceae). *Syst. Biol.* **57,** 466–482.

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.

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

Moore MJ, Bell CD, Soltis PS, Soltis DE (2007) Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *Proc. Natl. Acad. Sci. USA* **104,** 19363–19368.

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, Ramdial J, Alexandre R, Piedrahita A, Xi Z, Davis CC, Soltis PS, Soltis DE (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.

Mossel E,Vigoda E (2005) Phylogenetic MCMC algorithms are misleading on mixtures of trees. *Science* **309,** 2207–2209.

Murphy WJ, Pringle TH, Crider TA, Springer MS, Miller W (2007) Using genomic data to unravel the root of the placental mammal phylogeny. *Genome Res.* **17,** 413–421.

Oliver JC (2013) Microevolutionary processes generate phylogenomic discordance at ancient divergences. *Evolution* **67,** 1823–1830.

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 (1997) From gene to organismal phylogeny: Reconciled trees and the gene tree/species tree problem. *Mol. Phylogenet. Evol.* **7,** 231–240.

Page RDM, Charleston MA (1998) Trees within trees: Phylogeny and historical associations. *Trends Ecol. Evol.* **13,** 356–359.

Page RDM (2000) Extracting species trees from complex gene trees: Reconciled trees and vertebrate phylogeny. *Mol. Phylogenet. Evol.* **14,** 89–106.

Pamilo P, Nei M (1988) Relationships between gene trees and species trees. *Mol. Biol. Evol.* **5,** 568–583.

Pelser PB, Kennedy AH, Tepe EJ, Shidler JB, Nordenstam B, Kadereit JW, Watson LE (2010) Patterns and causes of incongruence between plastid and nuclear Senecioneae (Asteraceae) phylogenies. *Am. J. Bot.* **97,** 856–873.

Penny D, White WT, Hendy MD, Phillips MJ (2008) A bias in ML estimates of branch lengths in the presence of multiple signals. *Mol. Biol. Evol.* **25,** 239–242.

Peters JL, Zhuravlev Y, Fefelov I, Logie A, Omland KE (2007) Nuclear loci and coalescent methods support ancient hybridization as cause of mitochondrial paraphyly between gadwall and falcated duck (*Anas spp.*). *Evolution* **61,** 1992–2006.

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.

Pisani D (2004) Identifying and removing fast-evolving sites using compatibility analysis: An example from the arthropoda. *Syst. Biol.* **53,** 978–989.

Pollard DA, Iyer VN, Moses AM, Eisen MB (2006) Widespread discordance of gene trees with species tree in *Drosophila*: Evidence for incomplete lineage sorting. *PLoS Genet.* **2,** e173.

Posada D, Crandall KA (2001) Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. *Proc. Natl. Acad. Sci. USA* **98,** 13757–13762.

QiuYL, Lee JB, Bernasconi-Quadroni F, Soltis DE, Soltis PS, Zanis M, Zimmer EA, Chen ZD, Savolainenk V, Chase MW (1999) The earliest angiosperms: Evidence from mitochondrial, plastid and nuclear genomes. *Nature* **402,** 404–407.

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.

Raamsdonk LWV, Smiech MP, Sandbrink JM (1997) Introgression explains incongruence between nuclear and chloroplast DNA-based phylogenies in *Allium* section *Cepa*. *Bot. J. Linn. Soc.* **123,** 91–108.

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.

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, Sinervo B, Linder CR, Ungerer MC, Arias DM (1996a) Role of gene interactions in hybrid speciation: Evidence from ancient and experimental hybrids. *Science* **272,** 741–744.

Rieseberg LH, Whitton J, Linder CR (1996b) Molecular marker incongruence in plant hybrid zones and phylogenetic trees. *Acta Bot. Neerl.* **45,** 243–262.

Rieseberg LH (1997) Hybrid origins of plant species. *Ann. Rev. Ecol. Syst.* **28,** 359–389.

Rokas A, Williams BL, King N, Carroll SB (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425,** 798–804.

Rokas A, Carroll SB (2006) Bushes in the tree of life. *PLoS Biol.* **4,** e352.

Rosenberg NA, Nordberg M (2002) Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Rev.* **3,** 380–390.

Rosenberg NA (2003) The shapes of neutral gene genealogies in two species: Probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. *Evolution* **57,** 1465–1477.

Rosenberg NA, Degnana JH (2010) Coalescent histories for discordant gene trees and species trees. *Theor. Popul. Biol.* **77,** 145–151.

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.

Salichos L, Rokas A (2013) Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature*, doi: 10.1038/nature12130.

Sang T, Zhong Y (2000) Testing hybridization hypotheses based on incongruent gene trees. *Syst. Biol.* **49,** 422–434.

Savolainen V, Chase MW, Hoot SB, Morton CM, Soltis DE, Bayer C, Fay MF, De Bruijn AY, Sullivan S, Qiu YL (2000a) Phylogenetics of flowering plants based on combined analysis of plastid *atpB* and *rbcL* gene sequences. *Syst. Biol.* **49,** 306–362.

Savolainen V, Fay MF, Albach DC, Backlund A, van der Bank M, Cameron KM, Johnson SA, Lledó MD, Pintaud JC, Powell M, Sheahan MC, Soltis DE, Soltis PS, Weston P, Whitten WM, Wurdack KJ, Chase MW (2000b) Phylogeny of the eudicots: A nearly complete familial analysis based on *rbcL* gene sequences. *Kew Bull.* **55,** 257–309.

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

Scotland RW, Sanderson MJ (2004) The significance of few versus many in the tree of life. *Science* **303,** 643–643.

Seelanan T, Schnabel A, Wendel JF (1997) Congruence and consensus in the cotton tribe (Malvaceae). *Syst. Bot.* **22,** 259–290.

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, Burns P, Davis TM, Slovin JP, Bassil N, Hellens RP, Evans C, Harkins T, Kodira C, Desany B, Crasta OR, Jensen RV, Allan AC, Michael TP, Setuba JC, Celton J, Rees DJG, Williams KP, Holt SH, Rojas JJR, Chatterjee M, Liu B, Silva H, Meisel L, Adato A, Filichkin SA, Troggio M, Viola R, Ashman T, Wang H, Dharmawardhana P, Elser J, Raja R, Priest HD, Jr Bryant DW, Fox SE, Givan SA, Wilhelm LJ, Naithani S, Christoffels A, Salama DY, Carter J, Girona EL, Zdepski A, Wang W, Kerstetter RA, Schwab W, Korban SS, Davik J Monfort A, Denoyes-Rothan B, Arus P, Mittler R, Flinn B, Aharoni A, Bennetzen JL, Salzberg SL, Dickerman AW, Velasco R, Borodovsky M, Veilleux RE, Folta KM (2010) The genome of woodland

strawberry (*Fragaria vesca*). *Nat. Genet.* **43,** 109–116.

Simmons MP, Cappa JJ, Archer RH, Ford AJ, Eichstedt D, Clevinger CC (2008) Phylogeny of the Celastreae (Celastraceae) and the relationships of *Catha edulis* (qat) inferred from morphological characters and nuclear and plastid genes. *Mol. Phylogenet. Evol.* **48,** 745-757.

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.

Simpson MG (2012) Introgression, hybridization and polyploidy. In: *Plant systematics*, 2nd edn., Beijing, Science Press, pp. 653–655.

Slowinski JB, Page RDM (1999) How should species phylogenies be inferred from sequence data? *Syst. Biol.* **48,** 814–825.

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 DE, Soltis PS, Nickrent DL, Johnson LA., Hahn WJ, Hoot SB, Sweere JA, Kuzoff RK, Kron KA, Chase MW, Swensen SM, Zimmer EA, Chaw SM, Gillespie LJ, Kress WJ, Sytsma KJ (1997) Angiosperm phylogeny inferred from 18S ribosomal DNA sequences. *An. Mo. Bot. Gard.* **84,** 1–49.

Soltis DE, Soltis PS, Chase MW, Mort ME, Albach DC, Zanis M, Savolaninen V, Hahn WH, Hoot SB, Fay MF, Axtell M, Swensen SM, Prince LM, Kress WJ, Nixon KC, Farris J.S (2000) Angiosperm phylogeny inferred from 18S rDNA, *rbcL*, and *atpB* sequences. *Bot. J. Linn. Soc.* **133,** 381–461.

Soltis DE, Albert VA, Savolainen V, Hilu K, Qiu YL, Chase MW, Farris JS, Stefanović S, Rice DW, Palmer JD, Soltis PS (2004) Genome-scale data, angiosperm relationships, and "ending incongruence": a cautionary tale in phylogenetics. *Trends Plant Sci.* **9,** 477–83.

Soltis DE, Soltis PS (2004) *Amborella* not a "basal angiosperm"? Not so fast. *Am. J. Bot.* **91,** 997–1001.

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, Bell CD, Kim S, Soltis PS (2008) Origin and early evolution of angiosperms. *Ann. N. Y. Acad. Sci.* **1133,** 3–25.

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

Soltis DE, Burleigh G, Barbazuk WB, Moore MJ, Soltis PS (2009b) Advances in the use of next-generation sequence data in plant systematics and evolution. *ISHS Acta Horticulturae*,

859, 193–1206.

Soltis DE, Soltis PS (2009) The role of hybridization in plant speciation. *Annu. Rev. Plant Biol.* **60,** 561–588.

Soltis DE, Moore MJ, Burleigh JG, Bell CD, Soltis PS (2010) Assembling the angiosperm tree of life: progress and future prospects. *Ann. Mo. Bot. Gard.* **97,** 514–526.

Soltis DE, Smith SA, Cellinese N, Wurdack KJ, Tank DC, Brockington SF, Refulio-Rodriguez NF, Walker JB, Moore MJ, Carlsward BS, Bell CD, Latvis M, Crawley S, Black C, Diouf D, Xi Z, Rushworth CA, Gitzendanner MA, Sytsma KJ, Qiu YL, Hilu KW, Davis CC, Sanderson MJ, Beaman RS, Olmstead RG, Judd WS, Donoghue MJ, Soltis PS (2011) Angiosperm phylogeny: 17 genes, 640 taxa. *Am. J. Bot.* **98,** 704–730.

Soltis PS, Soltis DE (2013) Angiosperm phylogeny: A framework for studies of genome evolution. In: *Plant genome diversity* (eds. **Greilhuber J, Dolezel J, Wendel JF**). Vienna, Springer, Vol. 2, pp. 1-11.

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.

Stefanovic S, Rice DW, Palmer JD (2004) Long branch attraction, taxon sampling, and the earliest angiosperms: *Amborella* or monocots? *BMC Evol. Biol.* **4,** 35.

Stevens PF (2001 onwards) *Angiosperm phylogeny website.* Version 12, July 2012. http://www.mobot.org/MOBOT/research/APweb/.

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,** 1776–1782.

Wall, JD, Pritchard JK (2003) Assessing the performance of the haplotype block model of linkage disequilibrium. *Am. J. Hum. Genet.* **73,** 502–515.

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.

Wang N, Akey JM, Z Joly hang K, Chakraborty R, Jin L (2002) Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *Am. J. Hum. Genet.* **71,** 1227–1234.

Wehe A, Bansal MS, Burleigh JG, Eulenstein O (2008) DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. *Bioinformatics* **24,** 1540–1541.

Wendel JF, Doyle JJ (1998) Phylogenetic incongruence: Window into genome history and molecular evolution. In: *Molecular systematics of plants II: DNA sequencing* (eds. **Soltis DE,**

Soltis PS, Doyle JJ), Boston, Kluwer, pp. 265–296.

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.

Wortley AH, Rudall PJ, Harris DJ, Scotland RW (2005) How much data are needed to resolve a difficult phylogeny? Case study in Lamiales. *Syst. Biol.* **54,** 697–709.

Wurdack KJ, Davis CC (2009) Malpighiales phylogenetics: Gaining ground on one of the most recalcitrant clades in the angiosperm tree of life. *Am. J. Bot.* **96,** 1551–1570.

Xiang QYJ, Moody ML, Soltis DE, Fan CZ, Soltis PS (2002) Relationships within Cornales and circumscription of Cornaceae: *matK* and *rbcL* sequence data and effects of outgroups and long branches. *Mol. Phylogenet. Evol.* **24,** 35-57.

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.

Yang Z, Rannala B (2012) Molecular phylogenetics: principles and practice. *Nat. Rev. Genet.* **13,** 303–314.

Yoder JB, Briskine R, Mudge J, Farmer A, Paape T, Steele K, Weiblen GD, Bharti AK, Zhou P, May GD, Young ND, Tiffin P (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 J, Rowe WL, Struewing JP, Burtow KH (2002) HapScope: A software system for automated and visual analysis of functionally annotated haplotypes. *Nucleic Acids Res.* **30,** 5213–5221.

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.

Zeng LP, Zhang N, Ma H (2014) Advances and challenges in resolving the angiosperm phylogeny. *Biodiversity Science* **22,** 21–39.

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.

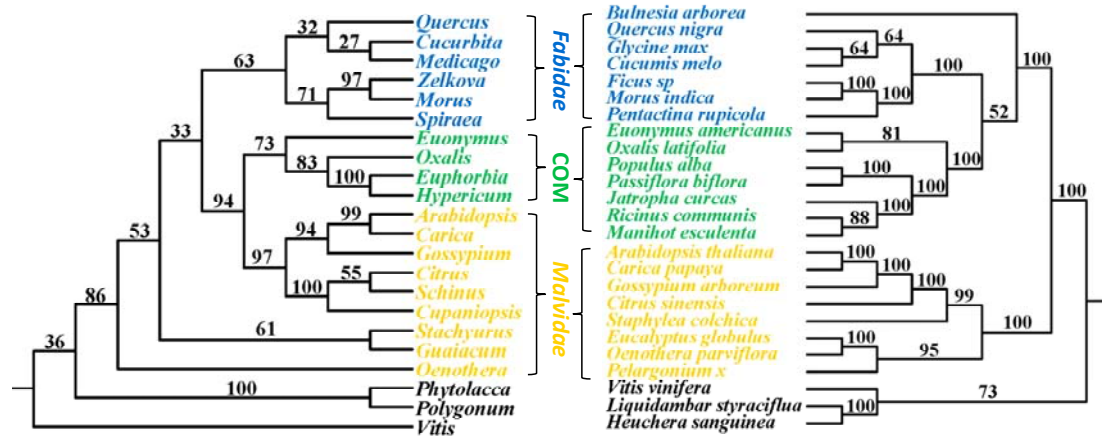
Zhu XY, Chase MW, Qiu YL, Kong HZ, Dilcher DL, Li JH, Chen ZD (2007) Mitochondrial *matR* sequences help to resolve deep phylogenetic relationships in rosids. *BMC Evol. Biol.* **7,**

217.

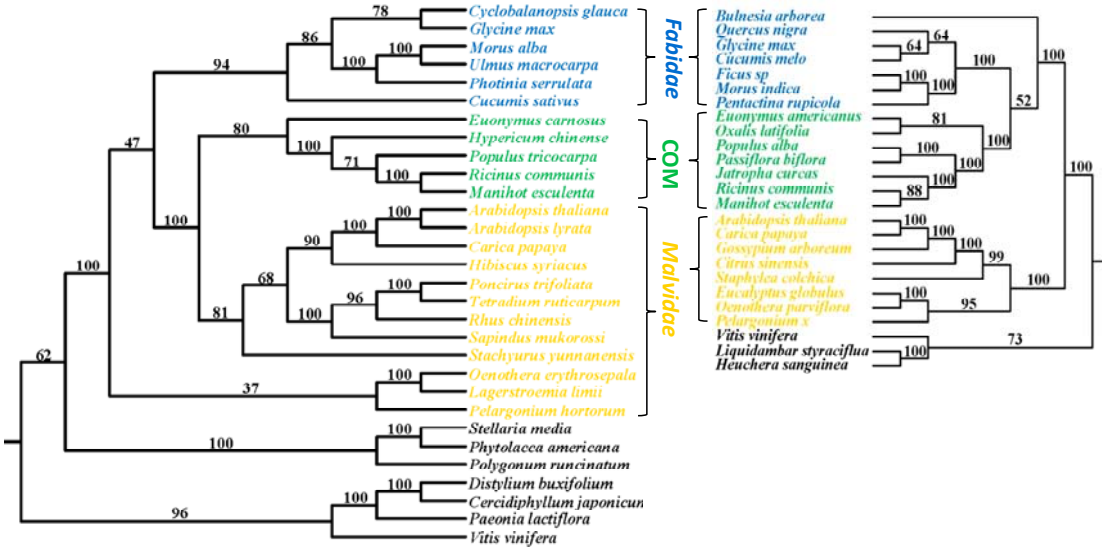
Zou XH, Ge S (2008) Conflicting gene trees and phylogenomics. *J. Syst. Evol.* **46,** 795–807.

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.

附录 A



Mitochordrial gene tree **Chloroplast gene tree**



Nuclear gene tree **Chloroplast gene tree**

图 A1 COM 支系统位置的叶绿体基因树分别与线粒体、核基因树冲突对比图

Figure A1 The comparison of phylogenetic positions of the COM clade inferred from Chloroplast, Mitochondrial, and Nuclear data sets

注：本图中的拓扑结构分别取自叶绿体、线粒体及核基因核苷酸矩阵 ML 树（图 3-1–图 3-3）；枝上的数字表示支持率。

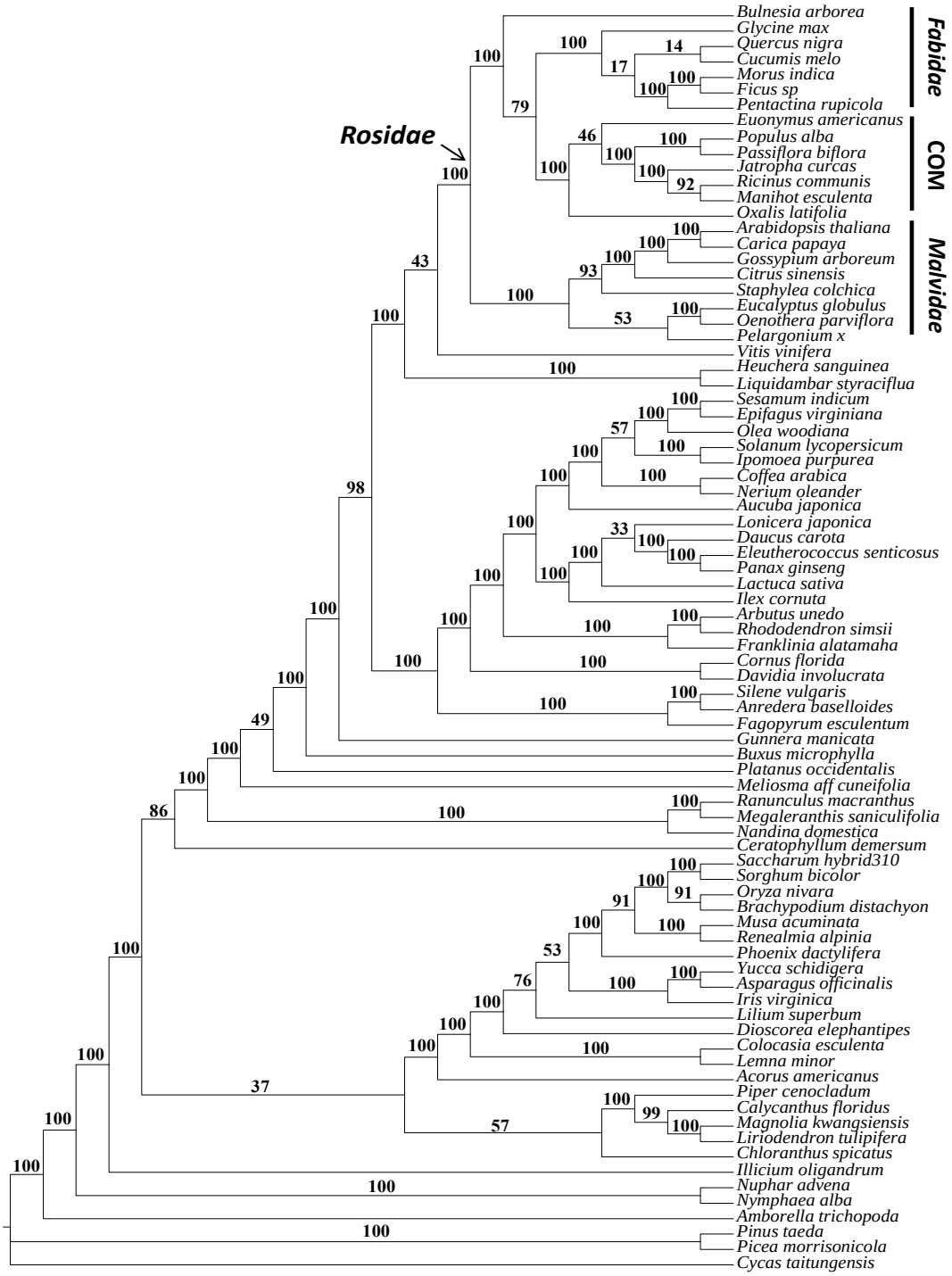


图 A2-1 叶绿体 78 基因 RY 编码矩阵多数一致 ML 树（枝上的数字表示支持率）

Figure A2-1 Maximum likelihood majority-rule bootstrap consensus tree inferred from the RY-coded (RY) matrix of 78 chloroplast genes (Numbers above the branches are Bootstrap value)

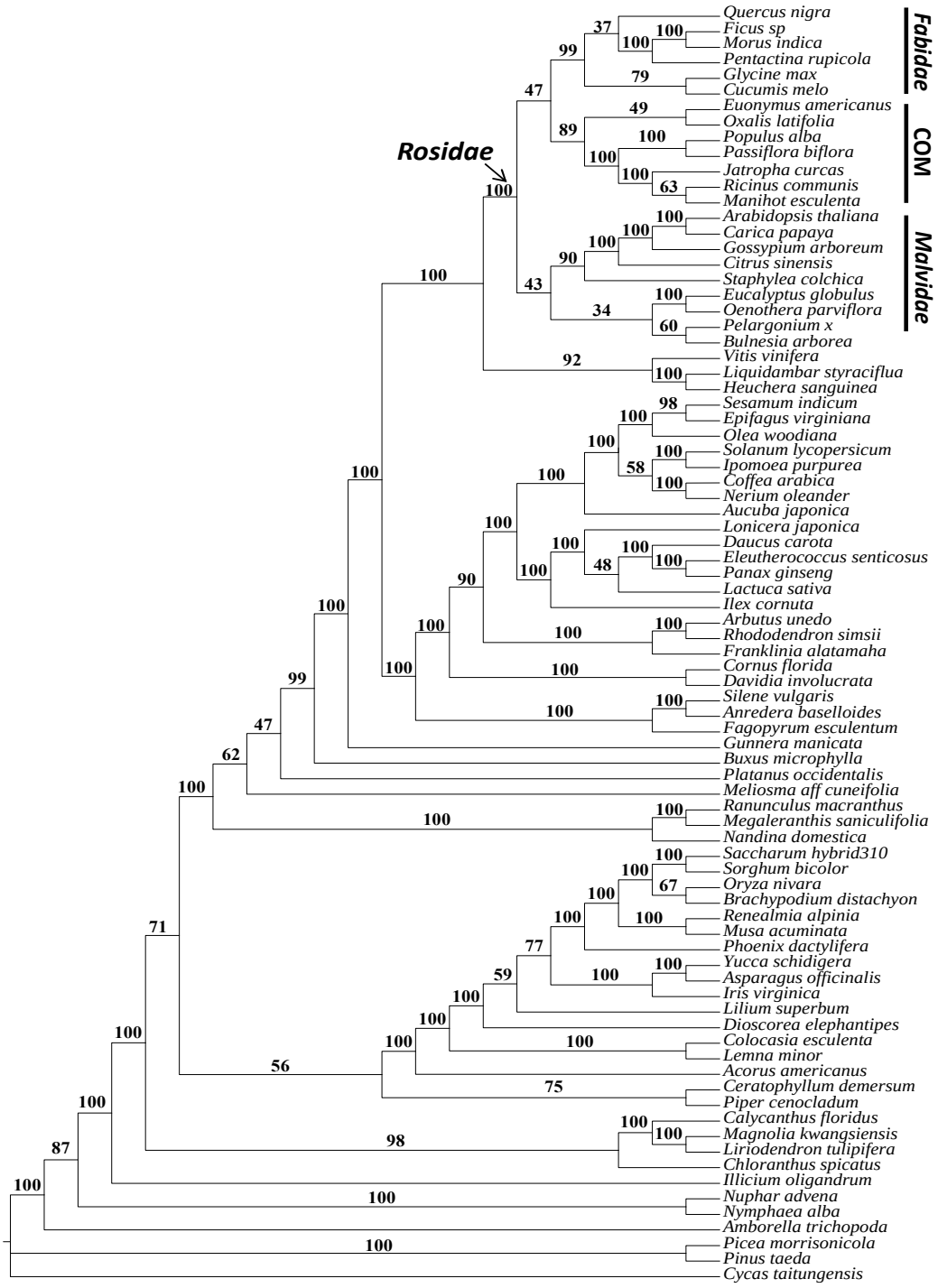


图 A2-2 叶绿体 78 基因氨基酸矩阵多数一致 ML 树

Figure A2-2 Maximum likelihood majority-rule bootstrap consensus tree inferred from the amino acid (AA) matrix of 78 chloroplast genes

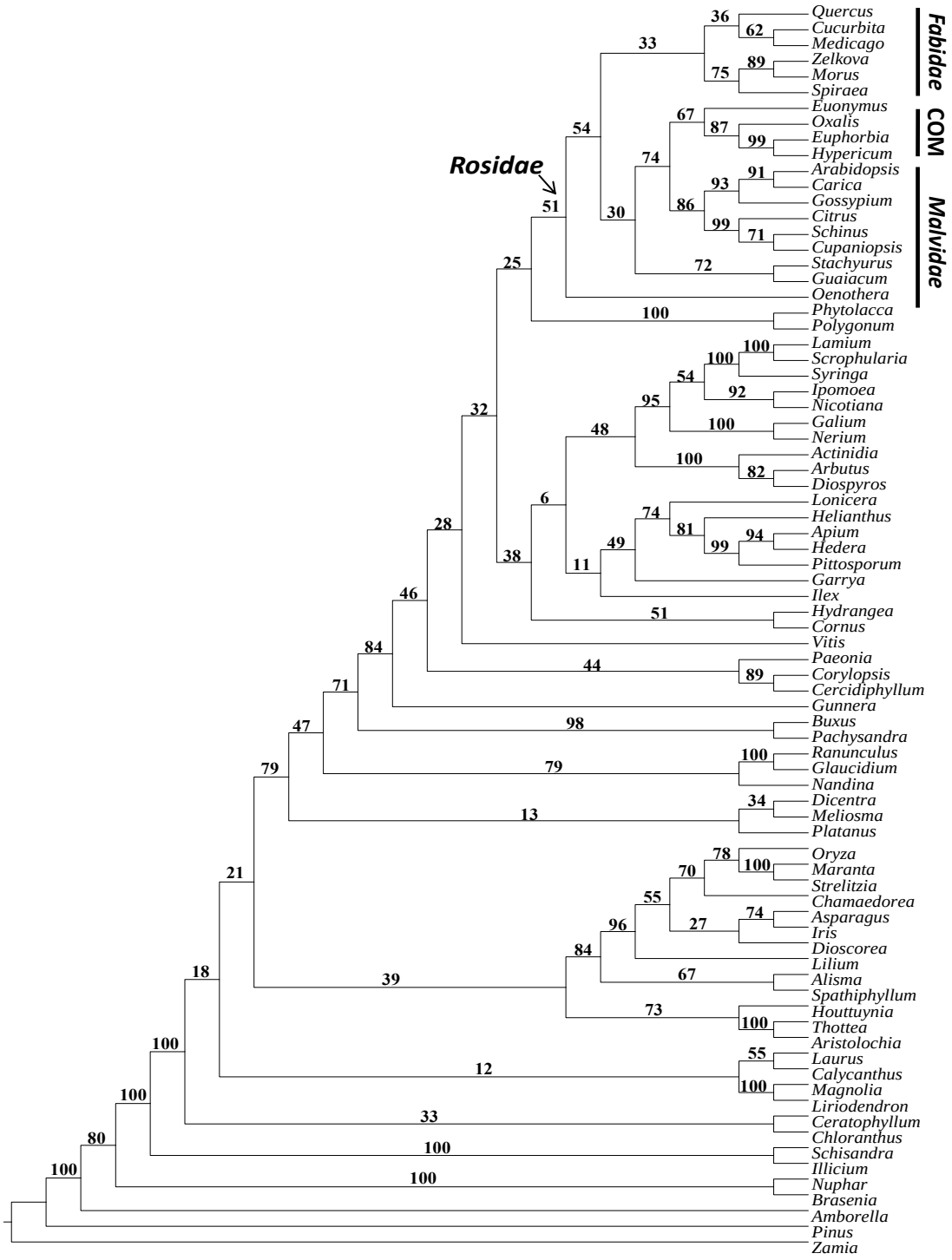


图 A3 线粒体 4 基因氨基酸矩阵多数一致 ML 树

Figure A3 Maximum likelihood majority-rule bootstrap consensus tree inferred from the amino acid (AA) matrix of 4 mitochondrial genes

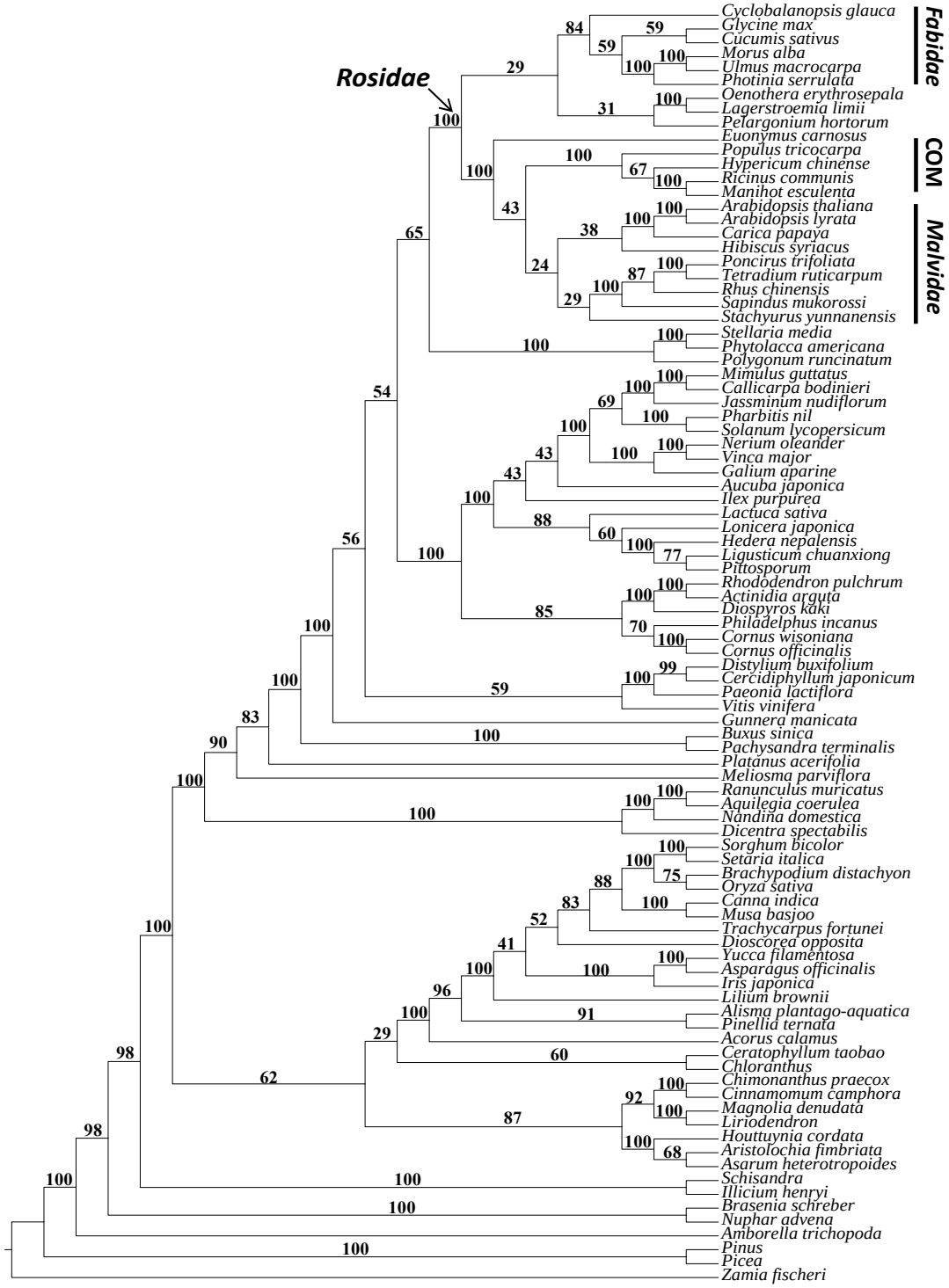


图 A4-1 核 5 基因 RY 编码矩阵多数一致 ML 树

Figure A4-1 Maximum likelihood majority-rule bootstrap consensus tree inferred from the RY coded (RY) matrix of 5 nuclear genes

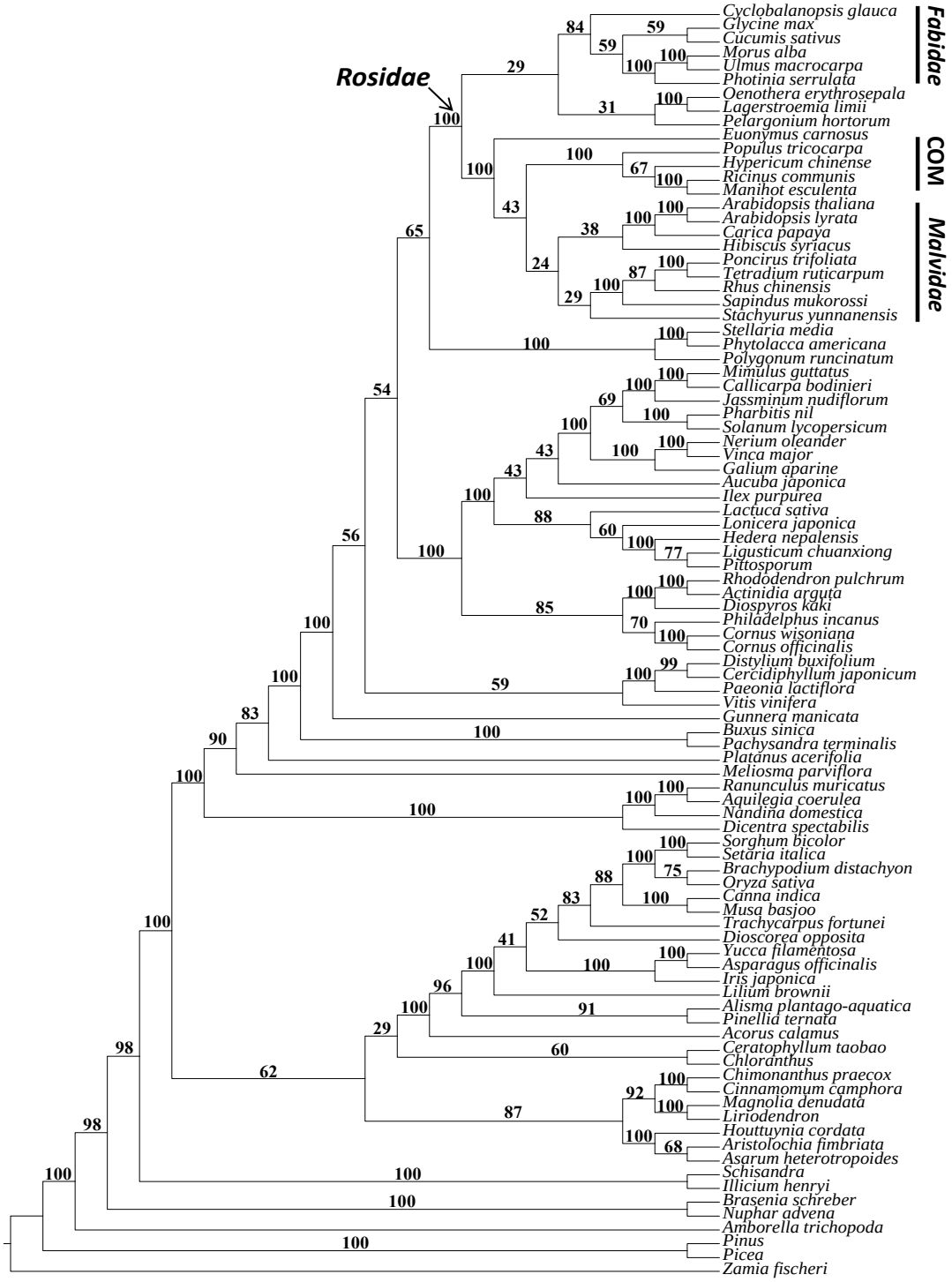


图 A4-2 核 5 基因氨基酸矩阵多数一致 ML 树

Figure A4-2 Maximum likelihood majority-rule bootstrap consensus tree inferred from the amino acid (AA) matrix of 5 nuclear genes

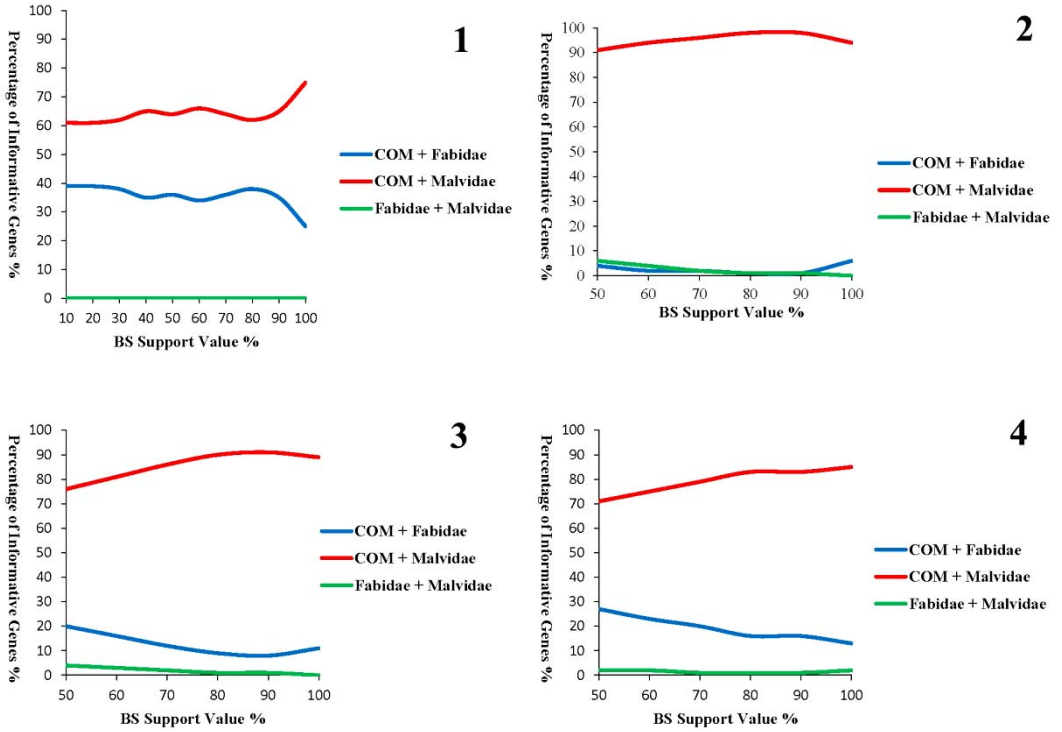


图 A5 核基因组中支持 COM 支三种拓扑关系对应基因数的百分比随支持率变化趋势图

Figure A5 Diagrams show the percentage of informative genes supporting the three putative hypotheses of the COM clade with the BS support increasing.

注：图 A5-1 单拷贝核基因分析中支持三种拓扑关系的基因数的百分比随支持率变化的趋势图（与表 3-2 对应）；图 A5-2–图 A5-4 多拷贝核基因分析中，依次在基因重复、基因重复与丢失以及不完全谱系筛选三种进化事件下，支持三种拓扑关系的基因数的百分比分别随支持率变化的趋势图（与表 3-3 对应）。

附录 B

叶绿体、线粒体和核基因组矩阵的单基因分析

Individual gene analyses of Chloroplast, Mitochondrial, and Nuclear matrices

基因组	基因	系统位置	BS 支持率
Chloroplast	*accD*	COM + *Fabidae*	16%
	atpA	COM + *Fabidae*	18%
	atpB	COM + *Fabidae*	17%
	atpE	COM + *Fabidae*	6%
	atpF	COM + *Malvidae*	3%
	atpH	Nr	/
	atpI	*Fabidae* + *Malvidae*	**64%**
	ccsA	COM + *Fabidae*	16%
	cemA	COM + *Fabidae*	47%
	clpP	COM + *Malvidae*	21%
	infA	Nr	/
	matK	COM + *Fabidae*	**93%**
	ndhA	COM + *Fabidae*	63%
	ndhB	COM + *Fabidae*	32%
	ndhC	Nr	/
	ndhD	COM + *Fabidae*	**83%**
	ndhE	Nr	/
	ndhF	COM + *Malvidae*	42%
	ndhG	Nr	/
	ndhH	COM + *Fabidae*	9%
	ndhI	Nr	/
	ndhJ	COM + *Fabidae*	29%
	ndhK	COM + *Fabidae*	2%
	petA	COM + *Fabidae*	35%
	petB	COM + *Malvidae*	13%
	petD	COM + *Fabidae*	8%
	petG	Nr	/
	petL	Nr	/
	petN	Nr	/
	psaA	COM + *Fabidae*	19%
	psaB	COM + *Fabidae*	39%
	psaC	Nr	/

psaI	Nr	/
psaJ	Nr	/
psbA	Nr	/
psbB	COM + *Fabidae*	10%
psbC	COM + *Fabidae*	32%
psbD	COM + *Fabidae*	14%
psbE	Nr	/
psbF	Nr	/
psbH	Nr	/
psbI	Nr	/
psbJ	Nr	/
psbK	Nr	/
psbL	Nr	/
psbM	Nr	/
psbN	Nr	/
psbT	Nr	/
psbZ	Nr	/
rbcL	COM + *Fabidae*	14%
rpl14	Nr	/
rpl16	COM + *Fabidae*	7%
rpl2	*Fabidae* + *Malvidae*	1%
rpl20	Nr	/
rpl22	COM + *Fabidae*	4%
rpl23	Nr	/
rpl32	COM + *Fabidae*	15%
rpl33	*Fabidae* + *Malvidae*	1%
rpl36	Nr	/
rpoA	COM + *Fabidae*	37%
rpoB	COM + *Fabidae*	35%
rpoC1	COM + *Fabidae*	**70%**
rpoC2	COM + *Fabidae*	40%
rps11	COM + *Fabidae*	1%
rps12	Nr	/
rps14	*Fabidae* + *Malvidae*	2%
rps15	COM + *Fabidae*	1%
rps16	*Fabidae* + *Malvidae*	8%
rps18	Nr	/
rps19	COM + *Fabidae*	1%
rps2	COM + *Fabidae*	38%
rps3	COM + *Malvidae*	25%

	rps4	COM + *Fabidae*	15%
	rps7	Nr	/
	rps8	Nr	/
	ycf2	COM + *Fabidae*	**91%**
	ycf3	COM + *Fabidae*	3%
	ycf4	COM + *Fabidae*	10%
Mitochordria	*atp1*	Nr	/
	nad5	Nr	/
	matR	COM + *Malvidae*	34%
	rps3	COM + *Malvidae*	51%
Nuclear	*SMC2*	COM + *Malvidae*	24%
	MCM5	COM + *Malvidae*	30%
	MSH1	COM + *Malvidae*	50%
	MLH1	COM + *Malvidae*	50%
	SMC1	COM + *Malvidae*	**80%**

注：a、Nr 表示对 COM 支没有支持，甚至整个蔷薇类都没有得到分辨；

b、表中加粗的数字表示 BS 支持率大于 60%。

致 谢

安排不如偶遇，偶遇不如巧合。我无意间读到了 Michael Donoghue 博士的一句话“*My research all revolves in one way or another around understanding phylogeny*”，顿感此语甚妙，耐人寻味，亦于我具颇强的反射性！我自本科北京林业大学到中国科学院植物所求学至今，专攻术业并未偏离过“植物学”这个关键词，且研究层级渐深，由表观的形态研究触及到亲缘进化的“*phylogeny*”。尽管些许辗转波折，然我甚是庆幸自己并未失此承接脉络。之所以如此，断不能离开诸位良师益友的如兰熏陶、教诲及慷慨帮助。毕业在即，感恩戴德！

我硕士师从本所标本馆林祁老师，学习胡颓子属（*Elaeagnus* L.）的分类学修订。期间，多次跟随马欣堂等标本馆考察队伍野外采集。此经历使我对植物分类浓生兴趣，大有“从一而终”之决心。但到硕士即将毕业准备博士深造之际，发现当时经典分类招生导师寥寥若无，现状倍感悲观，失望之至！正值彷徨迷失之际，承蒙马欣堂老师向我推荐从事被子植物系统发育研究且重视经典分类的陈之端老师。后来忆起他就是当时我在 PE 三楼查阅胡颓子属标本，而他在查阅葡萄科标本的那位。顿觉兴志有同而释怀，遂决然报考陈之端老师，拜其为师。始料未及的是当年报考者如水归海，竞争之烈，如惊涛骇浪，我终日忐忑，度日如年无须细表。当孔宏智老师电话于我告知被录取时，我和标本馆的杜玉芬老师高兴得竟然跳了起来！能够拜师之端门下，心想事成，实为人生难得美好之事！但每每思之，心存惭愧，总觉侥幸使然。总之，若无马欣堂老师之推荐，路安民老师、孔宏智老师、陈之端老师等诸位老师之认可，我断是不可能读到博士的。诸位老师拯救了我的理想，凭此一点，我就要对他们终生抱以感恩之情！

师从之端门下后，我自然没少得到尊师的启发、提点。恩师高瞻远瞩，儒雅博学，治学严谨却平易近人，以为人师表。在学期间，恩师一直对我寄予厚望，孜孜教诲，激励有加！故，在此首表感谢！在感激之余，为自己未能在学术上更求上进而惭愧，故我当继续奋进，以报知遇之恩！

同时，亦特别感谢路安民与王美林老师在生活和学业上给予我长辈关爱和温馨鼓舞！感谢朱新宇师兄、王伟师兄、李睿琦师姐、张剑师姐、以及组里同事曹志勇，还有杨拓、张景博、林立、李洪雷、武生聃、苏俊霞、胡瑾、王庆华、林若竹、向小果、鲁丽敏、董晓宇、陈闽、牛艳婷等同门之师兄弟姊妹，谢谢你们在学习和生活中给予的宽容、关爱和帮助！

感谢美国弗罗里达大学 Douglas E. Soltis、Pamela S. Soltis、Gordon J. Burleigh 博士在学术方面的毫无保留地指导和传授！再次感谢 Soltis 夫妇在访问期间给予我生活上热情的帮助，以及在海南、云南考察期间建立的忘年友谊！特别感谢留学生祁新帅、梅文彬等的热情帮助，同时与他们交流使我长了见识，宽了视

野，且省时有效，少走了不少弯路！

感谢中国科学院植物研究所系统与进化重点实验室这个优秀的集体为我创造的良好的学习和生活环境！感谢洪德元院士、葛颂研究员、汪小全研究员、孔宏智研究员、张宪春研究员、李良千研究员、贺超英研究员、朱相云研究员、李振宇研究员、罗毅波研究员、王印政研究员、贾渝研究员、周世良研究员、金效华副研究员、杨永副研究员、陈文俐副研究员、陈又生副研究员、王祺副研究员、邹新慧助理研究员、国春策助理研究员等老师们在我成长中的热情的帮助和指导！感谢冯旻老师、张宏耀老师、马欣堂老师、林祁老师、王忠涛老师、傅连中老师、杨志荣老师、班勤老师、杜玉芬老师、陈淑荣老师、田希娅老师、李爱莉老师、韦国芳老师、李敏老师、金健全老师、孟凡臣老师、靳婉青老师、梁荣花老师、乐寅婷老师和李哲老师的帮助，谢谢你们辛苦的工作为我完成学业创造了良好的条件！

感谢系统中心的王师傅，感谢植物园的王英伟、刘永刚老师，感谢外事处冷静、葛凤娟老师，感谢图书馆的韩芳桥老师，文献中心的刘凤红老师，感谢研究生部的赵宣、李蕊梅、张文娟、吴海燕、赵剑峰老师以及感谢研究生宿舍管理员王玉柱师傅，没有他们的种种的帮助，我也不可能顺利、安心地完成博士学业！

感谢中科院植物所足球队和所工会游泳俱乐部组织的各种活动，丰富并调节了我枯燥的科研和求学生活，同时也使我获得了健康的体魄和业余技能！

感谢刘冰、叶建飞、陈彬、卫然、张彩飞、赖阳均等对植物分类学兴趣浓厚的挚友们，曾多次有幸与你们一起进行野外考察和植物分类交流，受益颇多！更感谢在植物所期间学习和生活中给予我极大的支持、帮助、温暖和友谊的兄弟姊妹们，与你们朝夕相处的日子令我终生难忘！还有那些由于篇幅未列出的朋友、同学们，请原谅我的割爱，请接受我难于言表的感激！

最后，感谢家人和亲戚朋友们在生活上的理解、鼓励和支持！

2014 年 3 月 21 日于北京香山

个人简历

姓名：孙 苗　　出生年月：1983 年 2 月 15 日
性别：男　　籍　　贯：陕西省绥德县
专业：植物学

教育背景：

2002.9-2006.7	北京林业大学	环境科学专业（学士）
2006.9-2009.7	中国科学院植物研究所	植物学专业（硕士）
2009.9-2014.5	中国科学院植物研究所	植物学专业（博士）
2012.10-2013.1	*Florida Museum of Natural History, University of Florida*	访问学者

在学期间发表和待发表论文

Sun M, Naeem R, Su JX, Burleigh GJ, Solits DE, Soltis PS, Chen ZD (2014) Phylogeny of the *Rosidae*: A dense taxon sampling analysis. *Journal of Systematics and Evolution* (accepted).

Sun M, Solits DE, Soltis PS, Zhu XY, Burleigh GJ, Chen ZD (2014) Deep phylogenetic incongruence in the angiosperm clade *Rosidae* (submitted).

Lu LM, Sun M, Zhang JB, Li HL, Lin L, Yang T, Chen M, Chen ZD (2014) Tree of life and its applications. *Biodiversity Science* 22, 3-20.