



木犀科植物叶绿体基因组结构特征和系统发育关系

赵月梅¹, 杨振艳², 赵永平¹, 李筱玲¹, 赵志新¹, 赵桂仿^{3*}

¹商洛学院生物医药与食品工程学院, 商洛 726000; ²中国科学院昆明植物研究所, 东亚植物多样性与生物地理学重点实验室, 昆明 650201; ³西北大学生命科学学院, 西部资源生物与现代生物技术教育部重点实验室, 西安 710069

摘要 木犀科11属19个种叶绿体基因组的一般特征和变异特征的比较分析显示, 结果表明, 该科叶绿体基因组大小为154–165 kb, 其差异主要是大单拷贝(LSC)长度的差异所致。*Jasminum*属3个物种的叶绿体基因组长度与其余物种有较大差异, 该属*clpP*基因内含子和*accD*基因丢失。共线性分析表明, *Jasminum*属3个物种多个基因出现基因重排现象, 倒位可能是重排的主要原因。*Jasminum*属在IRb/SSC和SSC/IRa边界的基因均与其它物种不同; 重复序列与SSR数量检测结果表明, *Jasminum*属与其余物种在数量及重复长度上差异较大。基于CDS数据构建的系统发育树表明, *Abeliophyllum distichum*和*Forsythia suspensa*为木犀科中较早分化的类群。

关键词 木犀科, 叶绿体基因组, 结构特征, 系统发育

赵月梅, 杨振艳, 赵永平, 李筱玲, 赵志新, 赵桂仿 (2019). 木犀科植物叶绿体基因组结构特征和系统发育关系. 植物学报 54, 441–454.

叶绿体是绿色植物进行光合作用的细胞器, 具有合成蛋白质、淀粉及色素等功能, 其基因组可自主遗传, 在植物漫长的进化历史中起重要作用。在被子植物中, 多数叶绿体基因组(cpDNA)为母系遗传, 少数为双亲遗传或父系遗传(Hu et al., 2008)。叶绿体基因组结构和序列都比较保守, 大部分基因组结构为双链环状结构, 包括2个反向重复区(inverted repeats region, IRa/IRb)、1个大单拷贝区(large single copy region, LSC)和1个小单拷贝区(small single copy region, SSC) 4部分, 大小一般在120–180 kb之间(Zhang et al., 2012)。虽然与核基因组和线粒体基因组相比, cpDNA在基因组成和结构上相对保守, 但近年的研究也表明许多cpDNA发生基因变异, 如基因或内含子片段的丢失(Gao et al., 2010; 唐萍等, 2011)、反向重复区的长度变异或部分片段的插入/缺失(Chumley et al., 2006)、整个反向重复区的扩张或缺失(Chumley et al., 2006)及基因重排(Guisinger et al., 2011; Blazier et al., 2016)。由于基因组较小且结构相对保守, cpDNA已经成为进化和比较基因

组研究的理想模型(Dong et al., 2013), 为揭示植物类群的系统位置和遗传发育关系提供更多有利依据。

木犀科植物多为乔木, 直立或藤状灌木, 属于被子植物中合瓣花亚纲的木犀亚目, 包含28属400余种, 广泛分布于温带和热带, 我国有10属160余种(Flora of China Editorial Committee, 1996)。该科的茉莉(*Jasminum sambac*)、流苏(*Chionanthus retusus*)和迎春花(*J. nudiflorum*)等均是重要的园林植物; 连翘(*Forsythia suspensa*)则是重要的药用植物, 其果具清热解毒、消肿散结等功效。近年来, 随着分子系统学的发展, 有关木犀科的系统发育研究也有所增加(Wallander and Albert, 2000), 但多数研究仅基于核基因ITS或叶绿体基因片段, 且这些研究仅推断出木犀科部分属(如*Fraxinus*、*Syringa*、*Ligustrum*和*Olea*) (Jeandroz et al., 1997; Li et al., 2002; Wallander, 2008; Besnard et al., 2009)或种间的系统发育关系(Besnard et al., 2002; Baalicherif and Besnard, 2005)。随着二代测序技术的不断发展, 有关木犀科叶绿体基因组方面的研究也逐渐兴起。Lee

收稿日期: 2018-09-06; 接受日期: 2019-04-23

基金项目: 商洛市博士科技创新团队(No.SK2017-45)和商洛学院高层次人才引进项目(No.17SKY029)

* 通讯作者。E-mail: gfzhao@nwu.edu.cn

等(2007)主要分析了*Jasminum*和*Menodora*结构重排及其形成原因,内容上更侧重于叶绿体基因组的结构变异。时至今日,数据库中木犀科植物叶绿体基因组数据在不断更新,但尚未见对该科叶绿体基因组结构及进化关系的系统研究。本研究在数据库共享信息不断增加的基础上,拟通过下载整理数据库中木犀科19个物种的序列,从叶绿体基因组的角度分析这19个物种的种间关系,并通过比较这19个物种(11个属)的叶绿体基因组数据,揭示该科植物叶绿体基因组的结构特征和变异情况。

1 数据与方法

1.1 研究数据

从GenBank数据库下载已有的木犀科19个物种的叶绿体全基因组序列。物种名称、序列登录号及基因组具体信息见表1。

1.2 基因组统计和比较

利用Excel软件统计19个物种叶绿体基因组的大小、GC含量及LSC、SSC和IR区的长度与基因数目等信息。利用Geneious 8.0.2软件(<http://www.Geneious.com>)对每个序列的基因注释信息进行校对,并对部分注释进行手动修正,比较每个基因组各区域(LSC、SSC和IRs)的长度及边界信息;将每个叶绿体基因组序列的注释文件转换为bed格式,以*Abeliophyllum distichum*基因组作为参考序列,利用基因组在线分析程序mVISTA (<http://genome.lbl.gov/vista/mvista/submit.shtml>) (Frazer et al., 2004)中的Shuffle-Lagan模式进行全基因组比较;分别将每条基因组序列手动去除1个IR区,剩余部分导入Geneious 8.0.2软件中,利用Mauve插件(Darling et al., 2004)进行全局比对,以*A. distichum*基因组为参考序列,通过共线性分析检测基因组中的基因重排。

1.3 重复序列及SSR搜索

我们共识别木犀科19个物种叶绿体基因组中3种类型的重复序列,分别为串联序列(tandem repeat)、回文序列(panlindromic repeat)和散在序列(dispersed repeat)。其中串联重复用在线软件TRF (Tandem Repeats Finder) 查找(<http://tandem.bu.edu/trf/trf.html>) (Benson, 1999),重复序列长度设置为大于10 bp, Alignment Parameters中的匹配值、错配值和插入缺失分别设置为2、7和7,最小比对分值(Minimum Alignment Score)和最大比对长度(Maximum Period Size)分别设置为80和500。回文序列和散在序列使用在线分析软件REPuter统计(<https://bibiserv.cebitec.uni-bielefeld.de/reputer/>) (Kurtz et al., 2001)。参数设置:最小长度为30 bp (Minimal Repeat Size 30),最多碱基错配为3 (Hamming Distance 3)。利用MISA软件(<http://pgrc.ipk-gatersleben.de/misa/misa.html>)的Perl脚本查找叶绿体基因组序列中的简单重复序列(simple sequence repeats, SSRs) (Thiel et al., 2003),重复单元分别为单核苷酸(mononucleotide)、二核苷酸(dinucleotide)、三核苷酸(trinucleotide)、四核苷酸(tetranucleotide)、五核苷酸(pentanucleotide)和六核苷酸(hexnucleotide),重复次数阈值依次设定为10、5、4、3、3和3。

考虑本研究涉及的19个木犀科物种叶绿体基因组总长度及每个基因组各区域(LSC、SSC和IRs)的长度差异均较大,且*Jasminum*属中存在基因重组,我们选取19个物种的蛋白质编码基因(CDS)数据进行种间关系分析,删除在部分物种中缺失的基因。利用最大似然法(maximum likelihood, ML)和贝叶斯法(Bayesian inference, BI)构建系统发育树。在选择外类群时,我们根据GenBank中已有的数据信息,参考Lee等(2007)的研究结果,选择山茱萸科*Cornus controversa* (NC_030260)为外类群;同时,为保证结果的可靠性,我们还选择与木犀科同属于捩花目的夹竹桃科植物*Echites umbellatus* (NC_025655)和长春花(*Catharanthus roseus*) (NC_021423)为外类群(Flora of China Editorial Committee, 1995)。分别提取这3种植物叶绿体基因组的CDS区数据,与木犀科19种植物共同构建系统发育树。ML系统发育树用RAxML7.0.4软件(Stamatakis et al., 2008)生成,采用推荐的GTRGAMMA模型,bootstrap设置为1 000次。贝叶斯树(BI)在MrBayes 3.1.2软件中完成(Ronquist and Huelsenbeck, 2003),利用jModel-Test 2.1.4软件选择建树的最适模型(Posada, 2008),利用马尔可夫链(MCMC)算法,运行10 000 000代,

1.4 系统发育树构建及属间关系分析

考虑本研究涉及的19个木犀科物种叶绿体基因组总长度及每个基因组各区域(LSC、SSC和IRs)的长度差异均较大,且*Jasminum*属中存在基因重组,我们选取19个物种的蛋白质编码基因(CDS)数据进行种间关系分析,删除在部分物种中缺失的基因。利用最大似然法(maximum likelihood, ML)和贝叶斯法(Bayesian inference, BI)构建系统发育树。在选择外类群时,我们根据GenBank中已有的数据信息,参考Lee等(2007)的研究结果,选择山茱萸科*Cornus controversa* (NC_030260)为外类群;同时,为保证结果的可靠性,我们还选择与木犀科同属于捩花目的夹竹桃科植物*Echites umbellatus* (NC_025655)和长春花(*Catharanthus roseus*) (NC_021423)为外类群(Flora of China Editorial Committee, 1995)。分别提取这3种植物叶绿体基因组的CDS区数据,与木犀科19种植物共同构建系统发育树。ML系统发育树用RAxML7.0.4软件(Stamatakis et al., 2008)生成,采用推荐的GTRGAMMA模型,bootstrap设置为1 000次。贝叶斯树(BI)在MrBayes 3.1.2软件中完成(Ronquist and Huelsenbeck, 2003),利用jModel-Test 2.1.4软件选择建树的最适模型(Posada, 2008),利用马尔可夫链(MCMC)算法,运行10 000 000代,

表1 19个木犀科物种及其叶绿体基因组登录号和结构特征

Table 1 List of 19 species in Oleaceae, GenBank accession number and structural features of chloroplast genomes

No.	Species	References	GenBank No.	Genome size (bp)	LSC (bp)	SSC (bp)	IR (bp)	GC (%)	tRNA	rRNA	Protein-coding gene	Total gene	Pseudogene
1	<i>Abeliophyllum distichum</i>	Kim et al., 2016	NC_031445	155982	86742	17834	25703	37.8	37	8	88	134	1
2	<i>Chionanthus parkinsonii</i>	Van de Paer et al., 2018	NC_036979	155436	86230	17782	25712	37.8	37	8	88	133	
3	<i>C. retusus</i>	He et al., 2017	NC_035000	155687	86499	17894	25647	37.8	37	8	87	133	1
4	<i>C. rupicola</i>	Van de Paer et al., 2018	NC_036980	155563	86458	17813	25646	37.8	37	8	88	133	
5	<i>Fraxinus chiisanensis</i>	Kim et al., 2018	NC_037171	155541	86386	17767	25694	37.9	37	8	87	135	3
6	<i>F. excelsior</i>	Unpublished	NC_037446	155622	86415	17859	25674	37.8	37	8	88	133	
7	<i>Forestiera isabelae</i>	Van de Paer et al., 2018	NC_036981	155161	85921	17796	25722	37.9	37	8	88	133	
8	<i>Forsythia suspensa</i>	Wang et al., 2017	NC_036367	156404	87159	17811	25717	37.8	37	8	88	133	
9	<i>Jasminum nudiflorum</i>	Lee et al., 2007	NC_008407	165121	92877	13272	29486	38.0	38	8	87	133	
10	<i>J. sambac</i>	Unpublished	NC_034694	163186	90510	13252	29712	37.6	37	8	86	132	1
11	<i>J. tortuosum</i>	Unpublished	NC_034691	162080	89306	13280	29747	37.6	37	8	86	132	1
12	<i>Nestegis apetala</i>	Van de Paer et al., 2018	NC_036983	154849	85898	17505	25723	37.8	37	8	88	133	
13	<i>Noronia lowryi</i>	Van de Paer et al., 2018	NC_036984	155582	86410	17782	25695	37.9	37	8	88	133	
14	<i>Olea exasperata</i>	Van de Paer et al., 2018	NC_036985	155897	86687	17798	25706	37.8	37	8	88	133	
15	<i>O. europaea</i> subsp. <i>maroccana</i>	Gabriel et al., 2011	NC_015623	155896	86623	17789	25742	37.8	37	8	88	133	
16	<i>O. europaea</i> subsp. <i>europaea</i>	Gabriel et al., 2011	NC_015401	155875	86604	17787	25742	37.8	37	8	88	133	
17	<i>O. woodiana</i> subsp. <i>woodiana</i>	Gabriel et al., 2011	NC_015608	155942	86711	17799	25716	37.8	37	8	88	133	
18	<i>Schrebera arborea</i>	Van de Paer et al., 2018	NC_036986	155657	86622	17553	25741	37.8	37	8	88	133	
19	<i>Syringa vulgaris</i>	Van de Paer et al., 2018	NC_036987	155616	86242	17908	25733	37.9	37	8	88	133	

LSC: 大单拷贝区; SSC: 小单拷贝区; IR: 反向重复区 LSC: Large single copy region; SSC: Small single copy region; IR: Inverted repeats region

每1 000代取样1次。确定运行参数达到平衡后,舍弃25%的树作为老化样本(burn-in),用剩余样本生成一致树。

2 结果与分析

2.1 基因组结构和特征

叶绿体基因组结构包括大单拷贝区(LSC)、小单拷贝区(SSC)和2个反向重复区(IR)。本研究涉及的19个物种均未检测到IR大片段的缺失。但这些物种间叶绿体全基因组长度相差较大(表1),基因组长度最小的是 *Nestegis apetala*, 为154 849 bp, 最大的是 *Jasminum nudiflorum*, 长度为165 121 bp, 二者相差约10 kb。与其它物种类似,基因组4个区中,LSC区域最长,长度在85 898–92 877 bp之间,最小的是 *N. apetala*, 最大的是 *J. nudiflorum*, 二者相差约7 kb; SSC区长度在13 252–17 908 bp之间,相差约4.7 kb,最小的是 *J. sambac*, 最大的是 *Syringa vulgaris*; IR区长度为25 646–29 747 bp,相差约4 kb,最小的是 *Chionanthus rupicola*, 最大的是 *J. tortuosum*。15个基因(*rpl2*、*rpl16*、*rps16*、*petB*、*petD*、*ndhA*、*ndhB*、*rpoC1*、*atpF*、*trnG-GCC*、*trnL-UAA*、*trnK-UUU*、*trnA-UGC*、*trnI-GAU*和*trnV-UAC*)含有1个内含子,3个基因(*ycf3*、*clpP*和*rps12*)含有2个内含子。由此可知,LSC区长度的变异大于SSC区和IR区,19个物种的叶绿体基因组之间具有1–10 kb的差异,基因组长度的变异主要是LSC长度的差异引起;LSC区域和IR区域的最大值、SSC区的最小值均出现在 *Jasminum* 属,在总长度与分区长度上该属与木犀科其它属之间均差异较大。

2.2 基因数量

木犀科19个物种叶绿体总基因数为132–135个(表1),其中2个物种有132个基因,15个物种有133个基因,1个物种有134个基因,1个物种有135个基因。大部分基因为单拷贝,但位于IR区的7个蛋白编码基因(*rpl2*、*rpl23*、*rps7*、*ycf2*、*ndhB*、*ycf15*和*rps12*)、7个tRNA基因(*trnL-CAA*、*trnV-GAC*、*trnI-GAU*、*trnI-CAU*、*trnA-UGC*、*trnR-ACG*和*trnN-GUU*)和4个rRNA基因(*rrn16*、*rrn23*、*rrn4.5*和*rrn5*)均为2个拷贝。在3种类型的基因中,rRNA数量最保守,均为8个。

tRNA数目为37或38个,其中 *J. nudiflorum* 为38个,其余均为37个。蛋白编码基因为86–88个,其中 *accD* 基因在 *Jasminum* 属的3个种中丢失。另外,由于序列发布时间和注释方法不同,出现了在不同物种间同一序列注释不同的现象,因此以下基因名称混用: *ycf10* 和 *cemA*, *ycf5* 和 *ccsA*, *ycf9* 和 *psbZ*。

2.3 基因长度

在下载19个叶绿体基因组中,31个蛋白编码基因(包括5个含内含子基因和26个无内含子基因)长度有变异。其中含有内含子的基因有 *petD*、*rps12*、*clpP*、*rpoC1* 和 *rps16*,这5个基因中大部分基因编码区长度比较保守,基因长度差异主要取决于内含子的长度差异,甚至部分内含子丢失(表2)。*clpP* 基因除在物种

表2 叶绿体基因组中含内含子蛋白编码基因的编码区长度和基因全长

Table 2 Length of coding region and complete gene of intron-contained protein-coding genes of chloroplast genomes

Species	<i>petD</i>	<i>rps12</i>	<i>clpP</i>	<i>rpoC1</i>	<i>rps16</i>
1	483/1196	372/908	591/2047	2064/2821	267/1132
2	483/1218	372/908	591/2043	2073/2834	267/1147
3	483/1230	375/911	591/2045	2052/2820	267/1134
4	483/1261	381/917	591/2043	2073/2827	267/1153
5	483/1134	372/908	591/2045	2052/2811	255/1147
6	483/1148	372/908	591/2053	2073/2830	267/1136
7	483/1217	372/908	591/2039	2073/2834	267/1142
8	483/1196	372/908	588/2045	2064/2822	267/1131
9	483/1196	381/917	660	2076/2844	267/1162
10	483/1203	387/923	786	2052/2808	267/1154
11	483/1213	387/923	786	2052/2807	267/1161
12	483/1215	381/917	591/2041	2073/2833	267/1143
13	483/1217	372/908	591/2043	2073/2830	267/1143
14	483/1215	372/908	591/2046	2073/2832	267/1141
15	483/1213	371/907	591/2043	2073/2831	267/1142
16	483/1213	372/908	591/2044	2073/2831	267/1142
17	483/1215	373/909	591/2041	2073/2833	267/1141
18	483/1203	372/908	591/2047	2073/2834	267/1147
19	483/1199	372/913	591/2041	2073/2784	237/1115

物种编号1–19同表1。表中/前数字表示编码基因长度(bp),/后数字表示基因全长(bp)。

The species number 1–19 same as Table 1. The number before / indicates the length of coding region (bp), and the number after / indicates the full length of gene (bp).

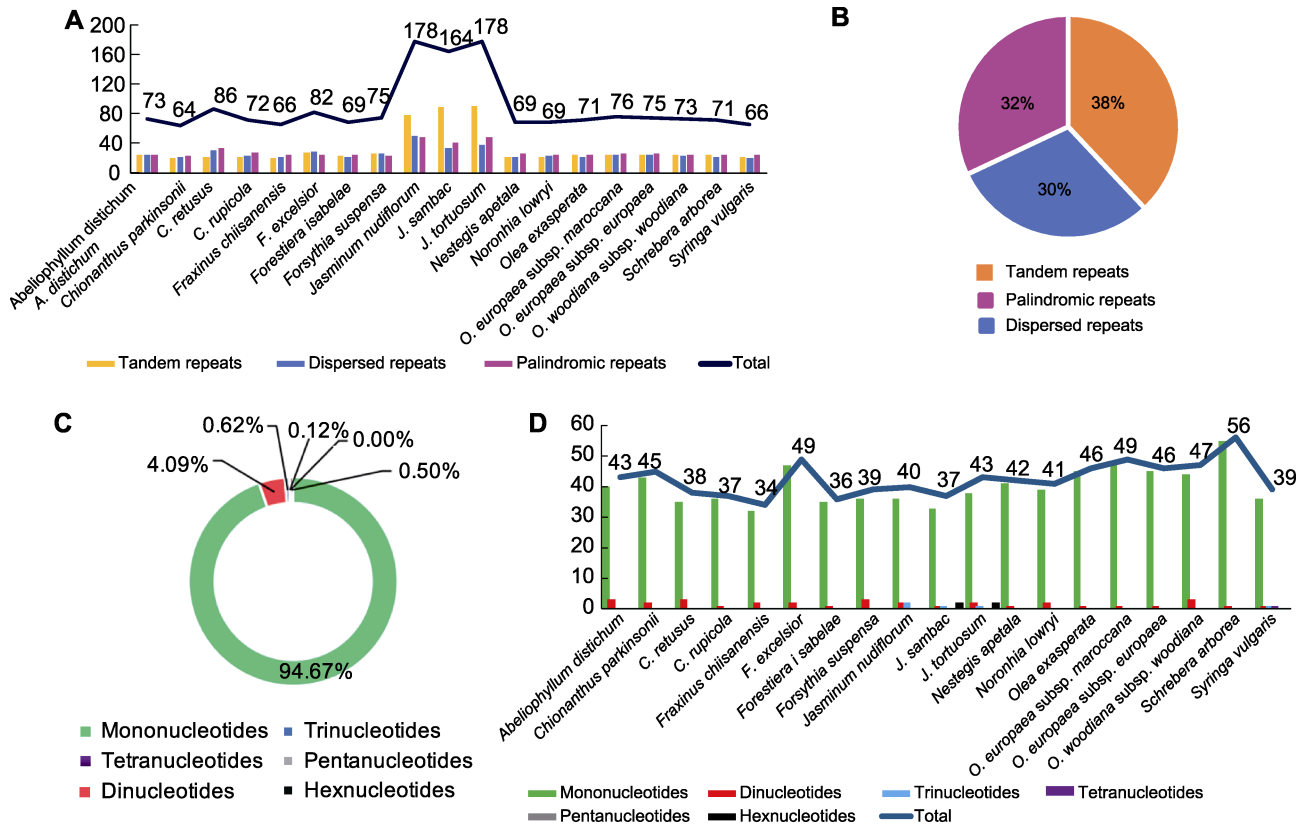


图1 19种木犀科植物叶绿体基因组中重复序列类型和比例
(A) 3种重复类型的数目; (B) 3种重复类型的比例; (C) SSR类型的比例; (D) SSR的数目和类型

Figure 1 The type and percentage of repeated sequences in the chloroplast genomes of 19 species in Oleaceae
(A) Number of three type repeats; (B) Percentage of three type repeats; (C) Percentage of SSR types; (D) Number and types of SSRs

*F. suspensa*中编码区长度存在差异外(588 bp), 在*Jasminum*属3个物种中均存在编码区长度改变和内含子丢失现象;*rpoC1*和*rps16*基因中未发现内含子片段缺失。与大部分物种相似,*rps12*基因存在反式剪切现象, 即该基因的5'端在LSC区, 3'端在IR区。无内含子的基因长度有差异(附表1)。不同基因差异程度不同, 如*atpA*和*psbH*只在1个物种中长度有差异,*rps15*、*ndhD*、*ccsA*、*rps19*、*rps3*、*infA*和*rps11*也只在少量物种中存在差异, 其余基因在不同物种中差异较大。

2.4 重复序列及SSR分析

在19种木犀科植物叶绿体基因组中我们共识别出1 677个重复序列, 包括502个散在重复、542个回文重复和633个串联重复(图1A, B; 附表2, 附表3)。其中,*Jasminum*属3个物种重复序列数量比其它物种明

显偏高, 分别为178 (*J. nudiflorum*)、164 (*J. sambac*)和178 (*J. tortuosum*)个, 该属3个物种的3种类型重复中, 串联序列的数量增加最明显(图1 A)。在重复序列长度方面, *Jasminum*属3个物种中均出现大量大于100 bp的重复, 其余16个物种均无此现象(附表3)。在19个木犀科物种中识别出807个简单重复序列(SSRs) (图1 C, D; 附表4)。在这些SSRs中, 最丰富的是单核苷酸SSR, 约占SSR总数的94.67% (764个), 二核苷酸SSR约占4.09% (33个), 三核苷酸SSR约占0.60% (5个), 四核苷酸SSR约占0.12% (1个), 六核苷酸SSR约占0.5% (4个), 这些物种中均未检测到五核苷酸SSR(图1C, D; 附表4)。其中三核苷酸重复分布在物种*J. nudiflorum* (2个)、*J. sambac* (1个)、*J. tortuosum* (1个)和*S. vulgaris* (1个)中; 四核苷酸重复分布在物种*S. vulgaris* (1个)中; 六核苷酸重复分布在*J. sambac* (2个)和*J. tortuosum* (2个)中。

2.5 边界分析

叶绿体基因组2个IR区域与LSC和SSC之间存在4个边界,即LSC/IRb、IRb/SSC、SSC/IRa和IRa/LSC。随着基因组进化,IR区域存在扩张与收缩现象,即有些基因会进入IR区或单拷贝区域。图2显示,木犀科19个物种叶绿体基因组中,*Jasminum*属3个物种与其它16个物种长度差异较大(29 486–29 747 bp),16个种IR长度均比较保守(25 646–25 742 bp)。本研究通过比较木犀科19个物种cpDNA的IR与SC边界区域基因的分布状况来推断木犀科物种在进化过程中发生的IR区域的扩张与收缩现象。如图2所示,LSC/IRb边界基因主要为*rps19*基因,但不同物种中该基因的位置有所差异,其中*A. distichum*、*Forestiera isabelae*和*S. vulgaris*这3个物种边界位于*rps19*基因内,3个种中*rps19*基因大部分位于LSC区,小部分位于IRb区,进入IRb区长度分别为2、1和2 bp。其它16个种*rps19*基因均位于LSC区,16个种中位于IRb区*rpl2*基因与LSC/IRb边界距离不等,介于15–62 bp之间;IRb/SSC边界在大部分物种中位于*ycf1*假基因和*ndhF*基因的重叠区内,重叠区域长度为23–99 bp。*Jasminum*属3个物种IRb/SSC边界中,*J. nudiflorum*和*J. tortuosum*中*ycf1*假基因和*ndhF*基因分别位于IRb区和SSC区,无重叠现象。*J. sambac*中IRb/SSC边界位于*ndhF*基因内,其中*ndhF*基因进入IRb区的长度为17 bp,该物种中*ycf1*假基因和*ndhF*基因无重叠,*ycf1*假基因只位于IRb区。除*Jasminum*属3个物种外其它种SSC/IRa边界均位于*ycf1*基因中,该基因在IRa区中的片段长度介于1 089–1 096 bp之间,但*Jasminum*属3个物种中*ycf1*基因均在IRa区域内,距离SSC/IRa边界133–148 bp。IRa/LSC与LSC/IRb相同。

2.6 全基因组水平比对及共线性分析

为了评估木犀科19个物种叶绿体基因组序列的差异程度,我们以*Abeliophyllum distichum*的序列为参考进行mVISTA分析(图3)。结果显示,除*Jasminum*属的3个种外,其余16个物种的序列高度相似。共线性分析(图4)显示,木犀科11个属19个种中,16个种的cpDNA序列均未检测到大片段的基因重排,说明相对保守。*Jasminum*属3个种均出现多个基因重排现象(如*ycf4*和*psaI*基因)。研究表明,这是该属在进化上的

重要特征(Lee et al., 2007)。此外,该属3个种还存在*accD*基因丢失和*clpP*基因内含子丢失现象。

2.7 系统发育分析

鉴于本研究所选用的木犀科这19个物种的叶绿体基因组序列总长度、LSC区、SSC区及IR长度差异较大且*Jasminum*属中的基因重组较多,我们选择这19个物种的叶绿体基因组共有的CDS数据构建系统发育树。以山茱萸科*Cornus controversa* (NC_030260)、夹竹桃科*Echites umbellatus* (NC_025655)和长春花(*Catharanthus roseus*) (NC_021423)为外类群,构建系统发育ML和BI树(图5),其中BI树使用的最佳核苷酸替换模型为GTR+I+G。结果显示,ML树和BI树拓扑结构一致,只有少数分支支持率稍有不同。木犀科11个属均各自聚为一支,且各分支支持率较高。其中*A. distichum*和*F. suspensa*亲缘关系较近,聚为一支,位于系统发育树的基部,为该科中较早分化的类群。*Jasminum*属3个物种聚在一起,形成独立的一支,位于*A. distichum*和*F. suspensa*分支的内部,该属分化也比较早。*Olea*属4个物种聚为一支,位于分支最内部,可能是木犀科中较晚分化出来的类群。

3 讨论

被子植物的叶绿体基因组大小为120–180 kb,IR区长度为20–30 kb (Zhang et al., 2012)。19个(11属)木犀科物种的叶绿体基因组大小为154–165 kb,IR区长25–29 kb,与被子植物叶绿体基因组长度相吻合。本研究表明,19个基因组大小差异约为10 kb,LSC、SSC与IR各区域的差异为LSC (7 kb)>SSC (4.7 kb)>IR (4 kb),可见木犀科不同种间基因组长度的差异主要是LSC的差异。研究表明,反向互补区可增强叶绿体基因组的稳定性和保守性(Khakhlova and Bock, 2006)。本研究中木犀科11属19种植物,除*Jasminum*属3个种外,其余16个种的叶绿体基因组IR区差异均很小,未发现IR区域过长、过短或丢失(Chumley et al., 2006; Hirao et al., 2008; Guisinger et al., 2011),整体较保守且一致。

叶绿体基因组IR区扩张与收缩是一种普遍的进化现象(Hansen et al., 2007; Huang et al., 2014)。例如,禾本目中由于LSC与IR边界扩张,*rps19*基因位于IR

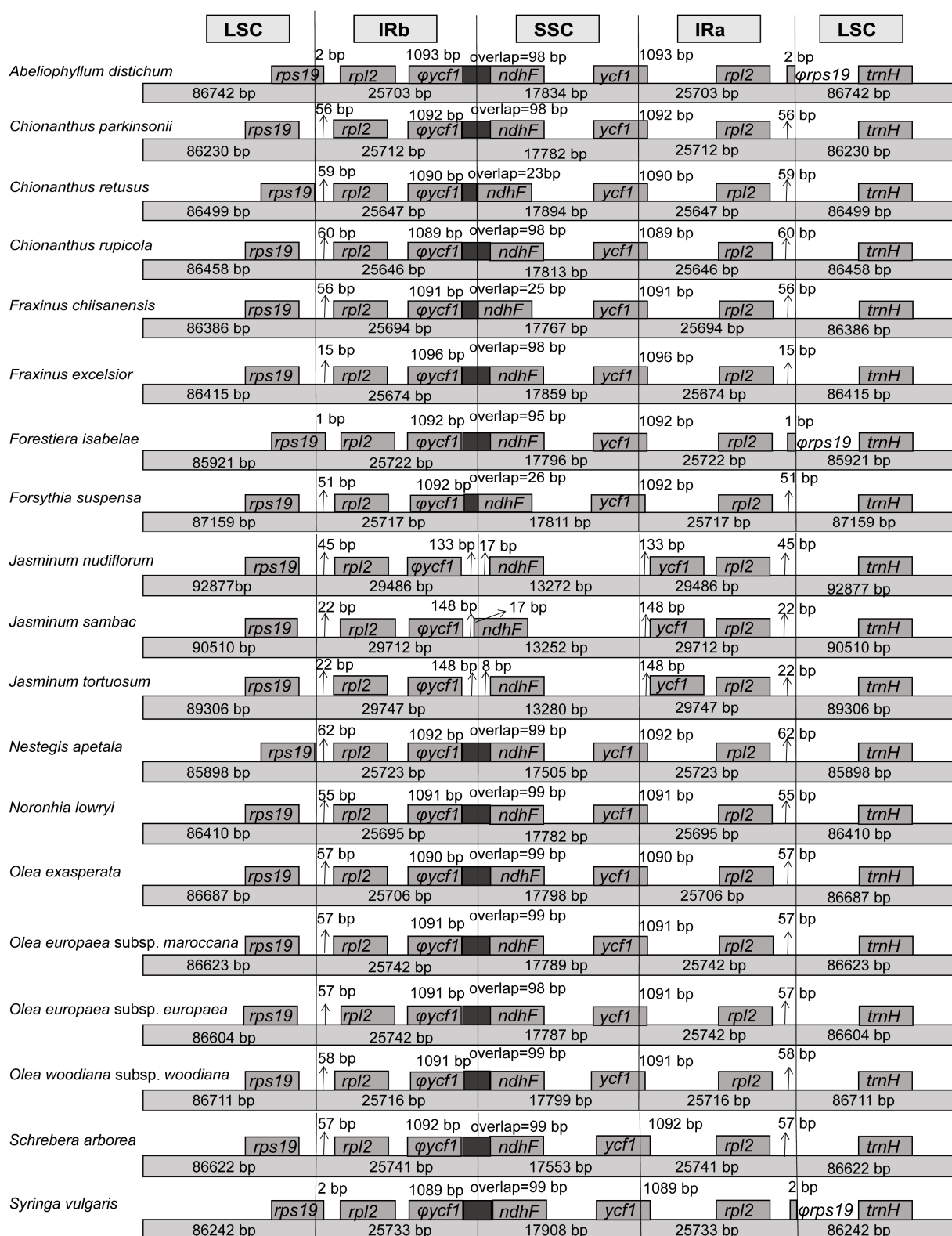


图2 木犀科19个物种的叶绿体基因组的四部分边界比较

Figure 2 Comparison of the border regions of the chloroplast genomes from 19 species in Oleaceae

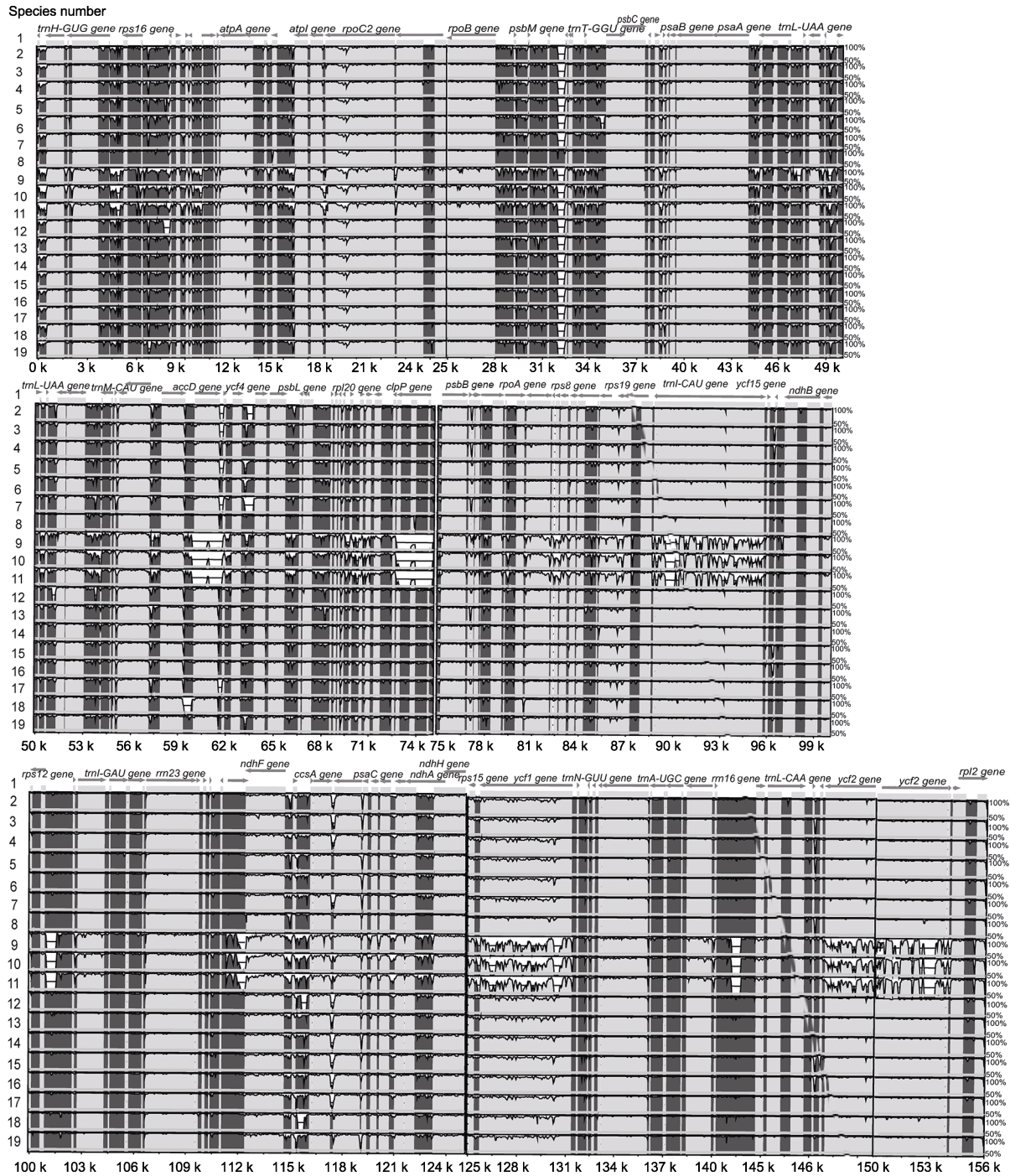


图3 木犀科19个物种叶绿体基因组的比对分析
物种编号1–19同表1。以*Abeliophyllum distichum*叶绿体基因组序列为参考序列(x轴), 各物种叶绿体基因组与参考序列一致度范围为50%–100%(y轴)。箭头表示基因及转录方向。

Figure 3 Graphic view of the alignment of chloroplast genomes from 19 species in Oleaceae
The species number 1–19 same as Table 1. Sequence identity varying between 50% and 100% are drawn on the y axis of the plot, the x axis corresponds to the coordinates on the *Abeliophyllum distichum* chloroplast genome. Arrows indicate the annotated genes and their transcriptional direction.

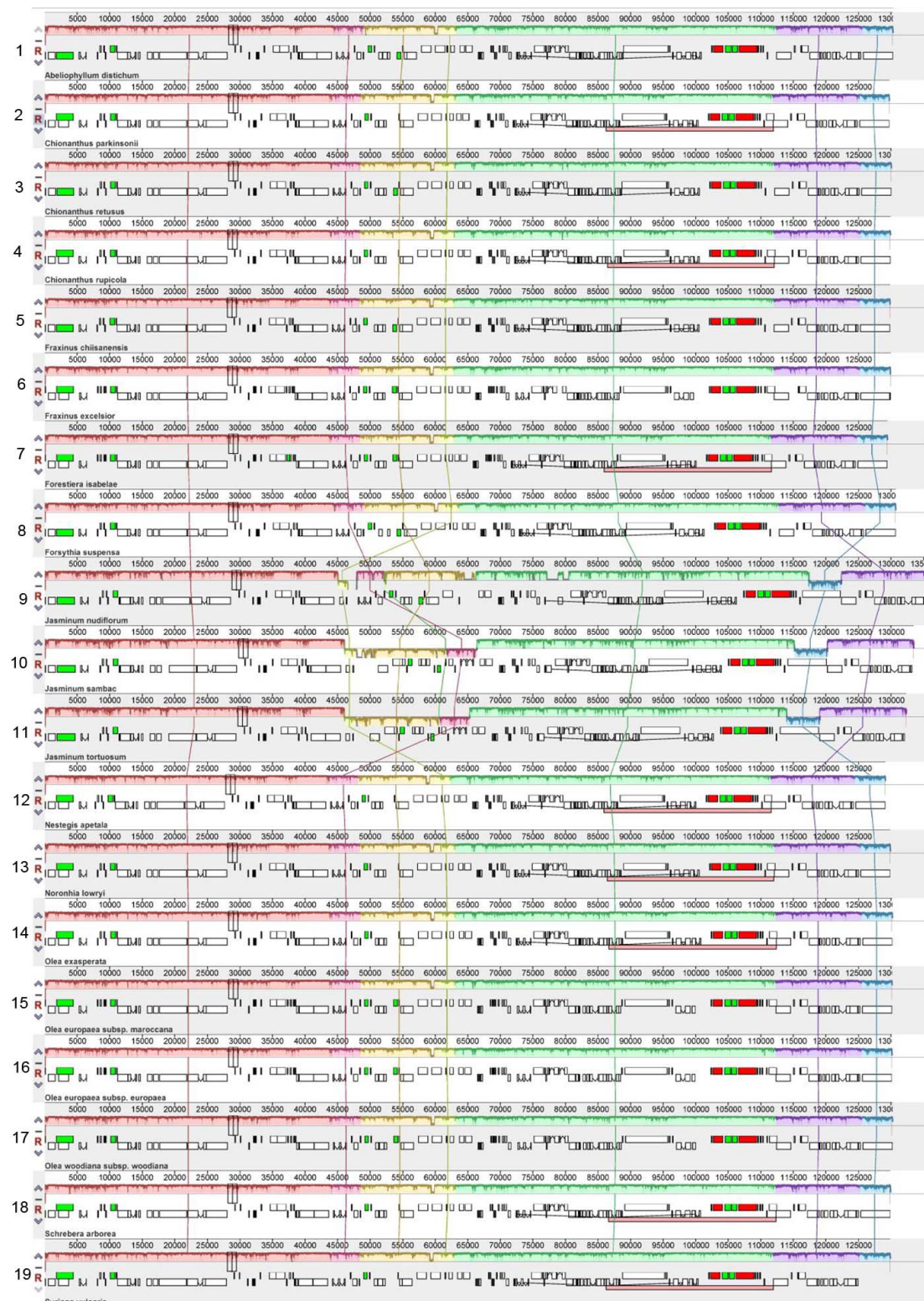


图4 木犀科19种植物的叶绿体基因组共线性分析
物种编号1-19同表1。

Figure 4 Synteny analyses of chloroplast genomes from 19 species in Oleaceae. The species number 1–19 same as Table 1.

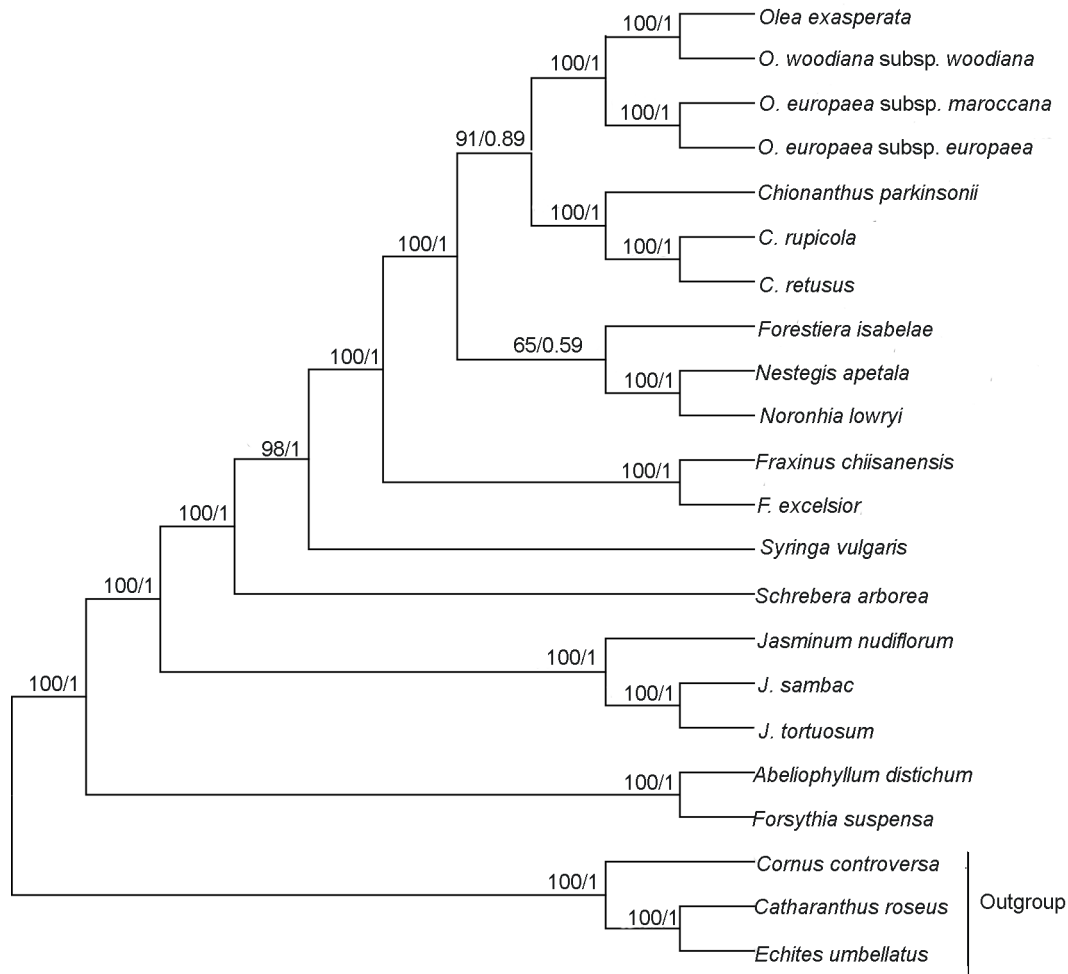


图5 基于木犀科19个物种共有蛋白编码基因构建的系统发育树
每个节点附近的数字分别表示ML树的自展支持率和BI树的后验概率。以*Cornus controversa*、*Echites umbellatus*和*Catharanthus roseus*为外类群。ML: 最大似然法; BI: 贝叶斯法

Figure 5 Phylogenetic relationship of 19 species in Oleaceae inferred from ML and BI analyses based on shared protein-coding genes
The numbers near each node are bootstrap support values in ML and posterior probability in BI. The outgroups are *Cornus controversa*, *Echites umbellatus* and *Catharanthus roseus*. ML: Maximum likelihood; BI: Bayesian inference

片段上, 但该类群中的大麦(*Hordeum vulgare*)、高粱(*Sorghum vulgare*)和剪股颖(*Agrostis matsumurae*)3种植物的IR区随后又发生了收缩, 使得*rps19*基因又进入LSC区中(唐萍等, 2011)。本研究中, 木犀科大部分物种的*rps19*基因位于LSC区, 少数物种的*rps19*基因(≤ 2 bp)进入IR区。大部分物种中IRb/SSC边界位于*ycf1*假基因和*ndhF*基因的重叠区内, 但*Jasminum*属3个物种IRb/SSC无基因重叠; SSC/IRa边界在除*Jasminum*属3个物种外均位于*ycf1*基因中, *Jasminum*属的3个物种中*ycf1*基因均在IRa区域内;

IRa/LSC边界在所有物种中均比较稳定。以上结果表明, *Jasminum*属物种边界基因与该科其它属物种差异较大。从长度上来看, *Jasminum*属整个基因组长度较该属其它物种长10 kb, 其中LSC和IR区域较长, SSC区域较短, 因此IRb/SSC和SSC/IRa边界与其余物种差异都很大, 即由于SSC区域较短, *ycf1*基因均进入邻近扩张的IR区域中。

叶绿体基因组序列共线性分析显示, 木犀科*Jasminum*属3个种的叶绿体基因组部分片段有重排现象。研究表明*Jasminum*属部分基因位置变化具有物

种特异性, 且可能与基因组结构有关(Lee et al., 2007; Greiner et al., 2008)。Lee等(2007)认为*J. nudiflorum*中的部分基因序列重排是由2段序列的先后倒位(inversion)所致, 该研究还标注了基因重排的具体位置并估算了倒位发生的大概时间。也有研究表明重复序列可能是倒位的原因之一(Palmer, 1985; Knox et al., 1993)。长且复杂的重复序列可能在基因组的重排或重组中起重要作用(Ogihara et al., 1988; Weng et al., 2014), 甚至重复序列之间的重组也可能诱导重排(Gray et al., 2009)。之前牻牛儿苗科发生了基因重排的天竺葵属和牻牛儿苗属中都含有大量的大片段重复序列(Guisinger et al., 2011)。本研究中出现基因重排的*Jasminum*属3种类型的重复序列无论数量还是长度均与其余物种有较大差异(图1A; 附表2, 附表3), 出现大量大于100 bp的重复序列, 数量也明显多, 与前人结论相符。在木犀科*Jasminum*属的3个物种中检测到的SSR数量与其余物种无明显差异, 但重复单元较长的3–6核苷酸重复大多分布在这3个物种中(附表4)。同其它植物类似(李巧丽等, 2018; Zhang et al., 2018), 本研究检测到的大量SSR也可作为潜在的分子标记用于该类群植物的后续研究。此外, 在多个物种中存在*accD*基因缺失也可能与基因重排有关(Hiratsuka et al., 1989; Maier et al., 1995; Katayama and Ogihara, 1996)。

本研究基于木犀科19个物种的叶绿体基因组共有的蛋白编码基因(CDS)数据构建系统发育树, 该科各属的关系均得到较好的处理, 11个属各自聚为一支且各节点均获得了较高的支持率。*A. distichum*和*F. suspensa*位于进化树的基部, 属木犀科中较为原始的类群。*Jasminum*中出现基因重排, 但基于保守的CDS序列构建的系统发育树表明该属并不是木犀科中的原始类群。*ndhF*和*rbcl*联合序列的系统发育分析也表明, *Abeliophyllum*、*Forsythia*、*Fontanesia*、*Myxopyrum*和*Nyctanthes*属为原始类群, 其分化起始于白垩纪早期(本研究中无后3个属的基因组数据), 支持*Jasminum*是单系类群且其分化起始于白垩纪晚期(78.3 MYA), 同时第1次DNA序列倒位也发生于该时期, 该属其它物种分化于始新世中期(42.1 MYA)(Lee et al., 2007)。Wallander和Albert (2000)曾用*rps16*和*trnL-F*序列对木犀科76个物种(25属)进行系统发育分析, 结果表明*Abeliophyllum*和*Forsythia*聚

在一起并位于进化树基部。本研究结果与其相符, 进一步在叶绿体基因组水平证明了这2个属在木犀科中的原始地位。但如上所述, 该文中*Fontanesia*和*Nactanthes*位于系统进化树最基部, 因数据缺乏, 我们无法对这2个研究中的系统发育关系进行更具体的比较。本研究中, *Olea*、*Nestegis*、*Chionanthus*、*Forestiera*、*Schrebera*、*Fraxinus*和*Syringa*属均为较晚出的分支, 但其具体的系统位置在不同序列构建的系统发育树中稍有不同(Wallander and Albert, 2000; Lee et al., 2007), 可能是所采用的数据集信息不同所致。此外, 本研究中样品数量不足也可能影响分析结果。

本研究基于目前已发表的叶绿体基因组数据, 分析了木犀科11个属19个物种的叶绿体全基因组结构和一般特征, 包括基因组大小、基因数量、长度、基因顺序和基因边界变化, 基于叶绿体基因组CDS数据构建的系统进化树各分支支持率均较高, 可为木犀科系统分类和系统发育关系研究提供有力的证据。随着高通量测序技术的不断发展, 系统发育关系研究已逐渐迈入基因组水平。然而, 相对于木犀科400多个物种而言, 当前的叶绿体基因组数量依然较少, 且由于该科部分属存在基因重组等问题, 未来仍需要利用更多的基因组数据对该科的属间和种间关系进行深入探讨。

参考文献

- 李巧丽, 延娜, 宋琼, 郭军战 (2018). 鲁桑叶绿体基因组序列及特征分析. 植物学报 53, 94–103.
- 唐萍, 阮秋燕, 彭程 (2011). 禾本科植物叶绿体基因组结构的系统进化研究. 中国农学通报 27(30), 17.
- Baali-Cherif D, Besnard G (2005). High genetic diversity and clonal growth in relict populations of *Olea europaea* subsp. *laperrinei* (Oleaceae) from Hoggar, Algeria. *Ann Bot* 96, 823–830.
- Benson G (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 27, 573–580.
- Besnard G, de Casas RR, Christin PA, Vargas P (2009). Phylogenetics of *Olea* (Oleaceae) based on plastid and nuclear ribosomal DNA sequences: tertiary climatic shifts and lineage differentiation times. *Ann Bot* 104, 143–160.
- Besnard G, Hernández P, Khadari B, Dorado G, Savolainen V (2011). Genomic profiling of plastid DNA varia-

- tion in the mediterranean olive tree. *BMC Plant Biol* **11**, 80.
- Besnard G, Khadari B, Baradat P, Bervillé A** (2002). *Olea europaea* (Oleaceae) phylogeography based on chloroplast DNA polymorphism. *Theor Appl Genet* **104**, 1353–1361.
- Blazier JC, Ruhlman TA, Weng ML, Rehman SK, Sabir JSM, Jansen RK** (2016). Divergence of RNA polymerase α subunits in angiosperm plastid genomes is mediated by genomic rearrangement. *Sci Rep* **6**, 24595.
- Chumley TW, Palmer JD, Mower JP, Fourcade HM, Calie PJ, Boore JL, Jansen RK** (2006). The complete chloroplast genome sequence of *Pelargonium x hortorum*: organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. *Mol Biol Evol* **23**, 2175–2190.
- Darling ACE, Mau B, Blattner FR, Perna ANT** (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* **14**, 1394–1403.
- Dong WP, Xu C, Cheng T, Lin K, Zhou SL** (2013). Sequencing angiosperm plastid genomes made easy: a complete set of universal primers and a case study on the phylogeny of *Saxifragales*. *Genome Biol Evol* **5**, 989–997.
- Flora of China Editorial Committee** (1995). Flora of China, Vol. 16. Beijing & St. Louis: Science Press & Missouri Botanical Garden Press. pp. 143–188.
- Flora of China Editorial Committee** (1996). Flora of China, Vol. 15. Beijing & St. Louis: Science Press & Missouri Botanical Garden Press. pp. 272–319.
- Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I** (2004). Vista: computational tools for comparative genomics. *Nucleic Acids Res* **32**, W273–W279.
- Gao L, Su YJ, Wang T** (2010). Plastid genome sequencing, comparative genomics, and phylogenomics: current status and prospects. *J Syst Evol* **48**, 77–93.
- Gray BN, Ahner BA, Hanson MR** (2009). Extensive homologous recombination between introduced and native regulatory plastid DNA elements in transplastomic plants. *Transgenic Res* **18**, 559–572.
- Greiner S, Wang X, Rauwolf U, Silber MV, Mayer K, Meurer J, Haberer G, Herrmann RG** (2008). The complete nucleotide sequences of the five genetically distinct plastid genomes of *Oenothera*, subsection *Oenothera*. I. Sequence evaluation and plastome evolution. *Nucleic Acids Res* **36**, 2366–2378.
- Guisinger MM, Kuehl JV, Boore JL, Jansen RK** (2011). Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and codon usage. *Mol Biol Evol* **28**, 583–600.
- Hansen DR, Dastidar SG, Cai ZQ, Penafior C, Kuehl JV, Boore JL, Jansen RK** (2007). Phylogenetic and evolutionary implications of complete chloroplast genome sequences of four early-diverging angiosperms: *Buxus* (Buxaceae), *Chloranthus* (Chloranthaceae), *Dioscorea* (Dioscoreaceae), and *Illicium* (Schisandraceae). *Mol Phylogenet Evol* **45**, 547–563.
- He YX, Liu LX, Yang SH, Dong MF, Yuan WJ, Shang FD** (2017). Characterization of the complete chloroplast genome of chinese fringetree (*Chionanthus retusus*). *Conserv Genet Resour* **9**, 431–434.
- Hirao T, Watanabe A, Kurita M, Kondo T, Takata K** (2008). Complete nucleotide sequence of the *Cryptomeria japonica* D. Don. chloroplast genome and comparative chloroplast genomics: diversified genomic structure of coniferous species. *BMC Plant Biol* **8**, 70.
- Hiratsuka J, Shimada H, Whittier R, Ishibashi T, Sakamoto M, Mori M, Kondo C, Honji Y, Sun CR, Meng BY, Li YQ, Kanno A, Nishizawa Y, Hirai A, Shinozaki K, Sugiura M** (1989). The complete sequence of the rice (*Oryza sativa*) chloroplast genome: intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. *Mol Gen Genet* **217**, 185–194.
- Hu YC, Zhang Q, Rao GY, Sodmergen** (2008). Occurrence of plastids in the sperm cells of Caprifoliaceae: biparental plastid inheritance in angiosperms is unilaterally derived from maternal inheritance. *Plant Cell Physiol* **49**, 958–968.
- Huang H, Shi C, Liu Y, Mao SY, Gao LZ** (2014). Thirteen *Camellia* chloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships. *BMC Evol Biol* **14**, 151.
- Jeandroz S, Roy A, Bousquet J** (1997). Phylogeny and phylogeography of the circumpolar genus *Fraxinus* (Oleaceae) based on internal transcribed spacer sequences of nuclear ribosomal DNA. *Mol Phylogenet Evol* **7**, 241–251.
- Katayama H, Ogihara Y** (1996). Phylogenetic affinities of the grasses to other monocots as revealed by molecular analysis of chloroplast DNA. *Curr Genet* **29**, 572–581.
- Khakhlova O, Bock R** (2006). Elimination of deleterious mutations in plastid genomes by gene conversion. *Plant J* **46**, 85–94.
- Kim C, Kim HJ, Do HDK, Jung J, Kim JH** (2018). Charac-

- terization of the complete chloroplast genome of *Fraxinus chliisanensis* (Oleaceae), an endemic to Korea. *Conserv Genet Resour* **11**, 63–66.
- Kim HW, Lee HL, Lee DK, Kim KJ** (2016). Complete plastid genome sequences of *Abeliophyllum distichum* Nakai (Oleaceae), a Korea endemic genus. *Mitochondr DNA Part B* **1**, 596–598.
- Knox EB, Downie SR, Palmer JD** (1993). Chloroplast genome rearrangements and the evolution of giant lobelias from herbaceous ancestors. *Mol Biol Evol* **10**, 414–430.
- Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R** (2001). Reputer: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res* **29**, 4633–4642.
- Lee HL, Jansen RK, Chumley TW, Kim KJ** (2007). Gene relocations within chloroplast genomes of *Jasminum* and *Menodora* (Oleaceae) are due to multiple, overlapping inversions. *Mol Biol Evol* **24**, 1161–1180.
- Li JH, Alexander JH, Zhang DL** (2002). Paraphyletic *Syringa* (Oleaceae): evidence from sequences of nuclear ribosomal DNA ITS and ETS regions. *Syst Bot* **27**, 592–597.
- Maier RM, Neckermann K, Igloi GL, Kösel H** (1995). Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. *J Mol Biol* **251**, 614–628.
- Ogihara Y, Terachi T, Sasakuma T** (1988). Intramolecular recombination of chloroplast genome mediated by short direct-repeat sequences in wheat species. *Proc Natl Acad Sci USA* **85**, 8573–8577.
- Palmer JD** (1985). Comparative organization of chloroplast genomes. *Annu Rev Genet* **19**, 325–354.
- Posada D** (2008). jModeltest: phylogenetic model averaging. *Mol Biol Evol* **25**, 1253–1256.
- Ronquist F, Huelsenbeck JP** (2003). Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574.
- Stamatakis A, Hoover P, Rougemont J** (2008). A rapid bootstrap algorithm for the raxml web servers. *Syst Biol* **57**, 758–771.
- Thiel T, Michalek W, Varshney R, Graner A** (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor Appl Genet* **106**, 411–422.
- Van de Paer C, Bouchez O, Besnard G** (2018). Prospects on the evolutionary mitogenomics of plants: a case study on the olive family (Oleaceae). *Mol Ecol Res* **18**, 407–423.
- Wallander E** (2008). Systematics of *Fraxinus* (Oleaceae) and evolution of dioecy. *Plant Syst Evol* **273**, 25–49.
- Wallander E, Albert VA** (2000). Phylogeny and classification of Oleaceae based on *rps16* and *trnL-F* sequence data. *Am J Bot* **87**, 1827–1841.
- Wang WB, Yu H, Wang JH, Lei WJ, Gao JH, Qiu XP, Wang JS** (2017). The complete chloroplast genome sequences of the medicinal plant *Forsythia suspensa* (Oleaceae). *Int J Mol Sci* **18**, 2288.
- Weng ML, Blazier JC, Govindu M, Jansen RK** (2014). Reconstruction of the ancestral plastid genome in Geraniaceae reveals a correlation between genome rearrangements, repeats, and nucleotide substitution rates. *Mol Biol Evol* **31**, 645–659.
- Zhang TW, Fang YJ, Wang XM, Deng X, Zhang XW, Hu SN, Yu J** (2012). The complete chloroplast and mitochondrial genome sequences of *Boea hygrometrica*: insights into the evolution of plant organellar genomes. *PLoS One* **7**, e30531.
- Zhang X, Zhou T, Yang J, Sun JJ, Ju MM, Zhao YM, Zhao GF** (2018). Comparative analyses of chloroplast genomes of Cucurbitaceae species: lights into selective pressures and phylogenetic relationships. *Molecules* **23**, 2165.

Chloroplast Genome Structural Characteristics and Phylogenetic Relationships of Oleaceae

Yuemei Zhao¹, Zhenyan Yang², Yongping Zhao¹, Xiaoling Li¹, Zhixin Zhao¹, Guifang Zhao^{3*}

¹College of Biopharmaceutical and Food Engineering, Shangluo University, Shangluo 726000, China; ²Key Laboratory for Plant Diversity and Biogeography of East Asia, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650201, China; ³Key Laboratory of Resource Biology and Biotechnology in Western China, Ministry of Education, College of Life Sciences, Northwest University, Xi'an 710069, China

Abstract In this study, the chloroplast genomes from 19 species (11 genera) in Oleaceae were compared to reveal the general characteristics and structural variations. The chloroplast genome sizes in Oleaceae were 154–165 kb, and the differences were mainly caused by the length of large single-copy regions. The chloroplast genome sizes of 3 species from the genus *Jasminum* differed greatly from that for other species; in addition, the introns from the *clpP* and *accD* genes were lost in *Jasminum*. Synteny analyses showed several gene rearrangements in 3 *Jasminum* species that were probably caused by inversions. The boundary genes between IRb/small single copy (SSC) and SSC/IRa regions in 3 *Jasminum* species differed from others. Repeat sequences and simple sequence repeat detection demonstrated that *Jasminum* had significant differences in repeat number and repeat length as compared with other genera. On the basis of shared protein-coding genes among 19 species, *Abeliophyllum distichum* and *Forsythia suspensa* were the early-diverging clades in Oleaceae.

Key words Oleaceae, chloroplast genomes, structural characteristics, phylogeny

Zhao YM, Yang ZY, Zhao YP, Li XL, Zhao ZX, Zhao GF (2019). Chloroplast genome structural characteristics and phylogenetic relationships of Oleaceae. *Chin Bull Bot* **54**, 441–454.

* Author for correspondence. E-mail: gzfzhao@nwu.edu.cn

(责任编辑: 白羽红)

附表 1 19 种木犀科植物叶绿体基因组中无内含子蛋白编码基因的长度

Appendix table 1 Length of protein-coding genes without introns in the chloroplast genome of 19 species in Oleaceae
<http://www.chinbullbotany.com/fileup/1674-3466/PDF/t18191-1.xlsx>

附表 2 19 种木犀科植物叶绿体基因组串联重复序列

Appendix table 2 List of tandem repeats in the chloroplast genome of 19 species in Oleaceae
<http://www.chinbullbotany.com/fileup/1674-3466/PDF/t18191-2.xlsx>

附表 3 19 种木犀科植物叶绿体基因组中的分散重复和回文重复序列

Appendix table 3 List of dispersed repeats and palindromic repeats in the chloroplast genome of 19 species in Oleaceae
<http://www.chinbullbotany.com/fileup/1674-3466/PDF/t18191-3.xlsx>

附表 4 19 种木犀科植物叶绿体基因组中简单重复序列(SSRs)信息

Appendix table 4 List of simple sequence repeats (SSRs) in the chloroplast genome of 19 species in Oleaceae
<http://www.chinbullbotany.com/fileup/1674-3466/PDF/t18191-4.xlsx>