植物学报 *Chinese Bulletin of Botany* 2023, **58** (2): 248–260, www.chinbullbotany.com
doi: 10.11983/CBB22026

·研究报告·

石栗叶绿体基因组研究

包金波，丁志杰，苗浩宇，李雪丽，任书贤，焦若岩，李浩
邓茜茜，李英姿，田新民*

新疆大学生命科学与技术学院，新疆生物资源基因工程重点实验室，乌鲁木齐 830017

摘要 石栗(*Aleurites moluccana*)是大戟科石栗属的常绿阔叶乔木，具有能源、药用和观赏价值。为填补石栗叶绿体基因组研究的空白，通过二代高通量全基因组测序，组装和注释了石栗叶绿体基因组，并进行基因组特征和系统发育分析。结果显示，石栗叶绿体基因组为典型的四段式结构，总长度为163 298 bp, LSC、SSC及IR的长度分别为91 301、18 501和26 748 bp。石栗叶绿体基因组共有131个基因，包括8个rRNA基因，37个tRNA基因，86个蛋白质编码基因。研究发现145个SSR位点，检测到重复单元有单核苷酸、二核苷酸、三核苷酸和四核苷酸，数目分别为80、53、10和2个。共线性分析结果表明，石栗叶绿体基因组存在基因倒位和重排现象。利用最大似然法和贝叶斯法构建了系统发育树，显示石栗与油桐(*Vernicia fordii*)和东京桐(*Deutzianthus tonkinensis*)亲缘关系较近，并形成姐妹群。利用化石时间进行定年分析，表明石栗属、油桐属和东京桐属的分化时间为25.94 Ma (95% HPD: 24.71–63.32 Ma)。该研究丰富了石栗基因组信息，可为石栗种质资源的开发利用提供基础遗传数据，同时为石栗属物种鉴定及系统发育研究提供参考。

关键词 石栗，叶绿体基因组，系统发育

包金波，丁志杰，苗浩宇，李雪丽，任书贤，焦若岩，李浩，邓茜茜，李英姿，田新民 (2023). 石栗叶绿体基因组研究. 植物学报 **58**, 248–260.

石栗(*Aleurites moluccana*)是大戟科石栗属的常绿阔叶乔木，耐旱，生长速度快，树高可达20 m。石栗原产于马来西亚及夏威夷群岛，分布于亚洲热带及亚热带地区，在我国分布于福建、台湾、广东、海南、广西及云南等省区。石栗为阳性树种，深根性，适生于肥沃湿润的酸性至中性土，喜高温高湿(梁文汇等, 2011)。石栗叶片长15–20 cm，果实为核果，具有木质纤维状坚硬果皮，果实内部含有1–3粒种子(图1)。种仁内粗蛋白(如蛋氨酸和赖氨酸)含量高达21.6%，粗脂肪含量高达66.22%，含油量高于大豆和菜籽(王磊, 2013)。此外，石栗油中富含不饱和脂肪酸亚麻酸和亚油酸(刘昌盛等, 2008)，活性成分分析表明其含有大量的鞣质、棕榈酸(曹晖等, 2007)以及甘油三酯(Radunz et al., 1998)。这些物化特性赋予石栗极其珍贵且特殊的应用价值。石栗是重要的绿化、能源和药用植物。作为绿化植物，石栗常被作为行道树广泛种植在道路两旁，是一种具有观赏价值的园艺树种，还可作为防风固沙的优势树种。作为能源植物，石栗油不仅可食用，经过纯化还可作为替代石油的生物柴油。研究表明，与未纯化的生物柴油相比，经深共晶溶剂(deep eutectic solvents, DESs)纯化的生物柴油的质量普遍得到改善，除了残炭和氧化稳定性外，均达到生物柴油可以接受的标准限值(Villarante and Ibarrientos, 2021)，是一种极具开发价值的木本油料能源树种。石栗的果壳具木质纤维状多孔结构，用其合成的活性炭复合材料作为生物吸附剂可吸附水溶液中的有毒铬离子(Vilarante et al., 2018)。作为药用植物，石栗果肉和果核可以做外敷止痛膏，树皮可用于治疗哮喘，花或果核制成活性炭可用于治疗咽喉肿痛(曹晖等, 2007)。石栗作为中草药产品，可有效治疗类风湿性关节炎和其它慢性疾病(Quintão et al., 2019)。在临床研究中，用0.5%和1.0%石栗提取物开发的制剂被证实是有效的镇痛剂、抗炎剂和伤口愈合剂(Cesca et al., 2012)。

收稿日期: 2022-02-10; 接受日期: 2022-05-10
基金项目: 国家自然科学基金(No.31601782)
* 通讯作者。E-mail: tianxm06@lzu.edu.cn

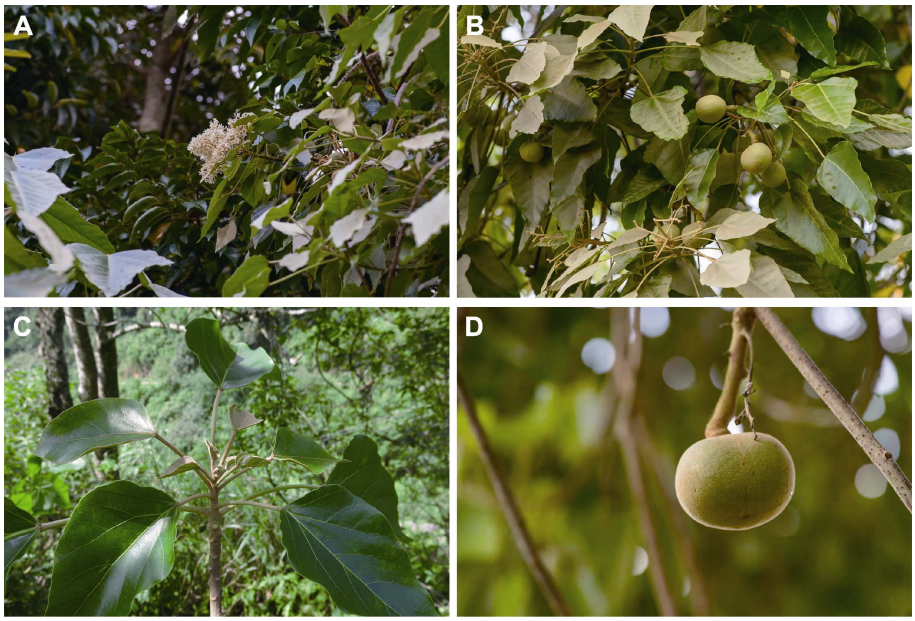

图1 石栗自然种群
(A) 花序; **(B)** 枝条; **(C)** 幼枝和幼叶; **(D)** 果实

Figure 1 Natural population of *Aleurites moluccana*
(A) Inflorescence; **(B)** Branches; **(C)** Young branches and leaves; **(D)** Fruit

长期以来, 对于石栗属物种的系统演化关系一直存在争议。1776年, Aleurites将石栗(*Aleurites moluccana* (L.) Willd)定为属, 即石栗属(*Aleurites* Forst)。而Joannis de Loureiro又新建立1个油桐属(*Vernicia* Lour)。1866年, Mneller-Argoviensis又将2属合为1属, 即油桐属(*Aleurites* Forst)。1966年, Airy Shaw将油桐属拆分为油桐属和石栗属(王劲风等, 1986)。由于与油桐形态特征相似, 石栗曾被划分为油桐的1个变种(蔡金标等, 1997), 但经过石栗与油桐属物种的叶绿体、氨基酸、DNA和RNA含量分析(苏梦云和周国璋, 1988)以及木材解剖结构比较(凌建群等, 1995)之后, 得出石栗属与油桐属存在显著差异。

叶绿体是高等植物细胞内活跃的代谢中心, 通过光合作用将太阳能转化为碳水化合物, 为细胞提供各种生命活动所需的能量(陈琴怡, 2017; 赵月梅等, 2019)。叶绿体还可发出调节细胞核中基因表达的信号, 即逆行信号(Krause, 2008), 在叶绿体调节次生代谢产物的合成方面具有重要作用。而叶绿体中合成的次生代谢物对于植物提高抗逆性和抵御病原菌非常重要。1986年, 烟草(*Nicotiana tabacum*)和地钱(*Marchantia polymorpha*)的叶绿体基因组测序数据发布, 这是叶绿体基因组首次被发现(李巧丽等, 2018)。植物叶绿体DNA一般为双链环状分子, 大小通常在120–170 kb之间。环状cpDNA包含反向重复区A (inverted repeat region A, IRA)和IRB、大单拷贝区(large single-copy region, LSC)与小单拷贝区(small single-copy region, SSC), 呈四段式结构(Shaw et al., 2007)。与核基因组相比, 叶绿体基因组较小, 便于测序; 结构也比核基因组更稳定; 叶绿体基因组序列高度保守, 遗传重组率低。基于上述优点, 叶绿体基因组全序列适合作为分子系统发育与分子生态学研究的遗传标记(Nie et al., 2012; Zhang et al., 2017)。

随着基因组测序技术的发展, 越来越多的叶绿体基因组被应用于系统进化研究领域(Daniell et al., 2016)。由于石栗具有重要的食用和药用价值, 目前对其研究多集中在化学成分活性、药效成分和影响产油量的叶绿体基因上, 而在系统发育、遗传多样性和叶绿体基因组方面研究较少。目前, 该物种的起源和演化、与其它石栗属物种的亲缘关系及其叶绿体基因组特征尚不清楚。

本研究利用二代高通量测序技术, 获得了石栗的

全基因组数据，从中组装和注释了叶绿体基因组，并结合已发表的大戟科物种叶绿体基因组数据，对石栗叶绿体基因组中简单重复序列(simple sequence repeat, SSR)、重复序列和密码子使用率等进行分析，重建了系统发育树，并确定该属的系统进化地位。本研究将为石栗的遗传保护及其在分子育种中的潜在应用提供参考。

1 材料与方法

1.1 实验材料及DNA的提取和测序

石栗(*Aleurites moluccana* (L.) Willd)个体采集于海南省热带植物园，标本号为2017-sl-001。使用植物DNA试剂盒(天根生物科技)提取基因组DNA。然后用琼脂糖凝胶电泳对DNA质量进行检测。将DNA经超声波破碎仪破碎后添加测序接头，破碎后的DNA小片段用PCR反应体系进行扩增，用CASpure PCR纯化试剂盒纯化。构建文库，建库类型为350 bp DNA小片段文库。构建好的文库经北京诺禾致源生物信息科技有限公司检测合格后，采用Illumina HiSeqPE150双端测序策略进行测序，测序深度为10倍。

1.2 序列组装及注释

基于Illumina HiSeq测序获得石栗全基因组序列，测序原始数据已上传至NCBI数据库，登录号为ON206671；并在SDB (Science Data Bank)数据库备份，登录号为10.57760/sciencedb.07992。为保证组装得到高质量叶绿体基因组序列，利用CLC Genomics Workbench v7.5软件对原始测序数据进行过滤，得到高质量clean reads。基于高质量数据，运用GetOrganelle v1.7.5 (https://github.com/Kinggerm/GetOrganelle)软件进行石栗叶绿体基因组的从头组装。首先将reads映射到seed，结合Seed Database组装seed reads并进行参数估计，然后通过迭代循环得到更多与靶序列相关的reads，接着进行从头组装，结合Label Database粗略过滤类似靶序列的contigs，同时使用SPAdes软件对clean reads进行从头组装，得到Cleaned and Labelled Target Assembly Graph；最后识别目标contigs并导出所有结构，得到组装结果，同时用Bandage v0.8.1 (http://rrwick.github.io/Bandage.)软件对从头组装得到的基因组结构图进行可视化，以确定最终组装结果。主要组装命令为：(get_organelle_from_reads. py-Fembplant_pt-t 1-R 15-o out_1-w 0.5-k 105,115,125,127)。

在基因注释软件GeSeq (https://chlorobox.mpimp-golm.mpg.de/geseq.html)中打开fasta文件，然后选中annotate plastid IR和annotate plastid trans-spliced *rps12*，将blast identity的阈值设置为85。为确保准确性，经过注释后的基因用DOGMA手动确认。

1.3 叶绿体基因组特征分析

利用MISA (http://pgrc.ipk-gatersleben.de/misa/)软件检测石栗叶绿体基因组序列中潜在的微卫星序列。用在线软件Reputer (https://bibiserv.cebitec.uni-bielefeld.de/repuer)统计石栗叶绿体基因组中重复序列的类型和数量，最大重复长度(maximum computed repeats)、最小重复长度(minimal repeat size)和Harming距离分别设置为50、30和3。使用Codon W1.4.4软件(linux版本)对叶绿体基因组序列中的氨基酸密码子偏好性进行统计，在Condon W初始页面中选择codon usage indic，设置密码子使用参数，选择select all，上传数据并输入Run c-condons运行程序，所有设置均为默认参数。用Mauve进行共线性分析，上传共线性分析所用的所有fasta序列，运行后输出结果文件，检测石栗在进化过程中基因是否发生位置变化。使用IRscope (https://irscope.shinyapps.io/irapp/)在线软件分析石栗叶绿体基因组IR区域的扩张与收缩现象，上传所有GB文件，点击submit等待输出结果。

1.4 系统发育分析

本研究选取大戟科物种包括油桐(*Vernicia fordii*)、东京桐(*Deutzianthus tonkinensis*)、木油桐(*V. montana*)、巴西橡胶树(*Hevea brasiliensis*)、少花橡胶树(*H. pauciflora*)、矮生小叶橡胶(*H. camargoana*)、木薯(*Manihot esculenta*)、麻风树(*Jatropha curcas*)，以及十字花科芥菜(*Brassica juncea*)、蔷薇科沙梨(*Pyrus pyrifolia*)、锦葵科刺槐棉(*Gossypium thurberi*)、芸香科来檬(*Citrus* × *aurantiifolia*)、葡萄科葡萄(*Vitis vinifera*)、铁青树科赤苍藤(*Erythropalum scandens*)、杨柳科南非醋栗(*Dovyalis caffra*)、大叶刺篱木(*Flacourtia rukam*)、菊科(*Scolopia chinen-*

sis)、广东箣柊(*S. saeva*)和红厚壳科铁力木(*Mesua ferrea*)。以上述物种的蛋白编码序列作为数据集，运用Phylosuite v1.2.1软件，利用最大似然(maximum likelihoods, ML)和贝叶斯(Bayesian, BI)法构建系统发育树。将用于构建系统进化树的每个物种的所有蛋白编码序列整合在fasta文件中，使用在线软件MAFFT对20条蛋白编码序列进行比对，比对结果在MEGA 7中选择Clustal W对齐，然后删除两端未对齐的碱基，将比对结果保存为fasta格式，并分别使用IQ-Tree和MrBayes进行ML和BI分析。使用内置程序Modelfinder设置模型，然后用IQ插件构建ML系统发育树。bootstrap值设为20 000，最优模型为GTR+F+R3，其它参数为默认，计算并生成ML系统发育树。用MrBayes插件构建BI树, MCMC (Markov Chain Montechains)为10 000 000代，抽样频率为1 000，最优模型为GTR+F+I+G4，运行过程中Average standard deviation of split frequencies值大于0.01时，需要增加MCMC代数，直至该值小于0.01，结果收敛，结束运行。

1.5 年龄估计与化石定年

将序列比对结果文件用Phylosuite软件转换成nex格式，利用BEAST v1.8.4软件中的BEAUti设置参数，设置Site Model参数，根据Phylosuite v1.2.1软件的Model Finder插件得到最优模型GTR。然后设置分子钟模型，选择松弛分子钟模型Relaxed clock log Normal，参数为默认。选择Yule tree prior为树先验模型，由于石栗属没有化石，选择白垩纪晚期最早的大戟科果实(66 Ma) (Reback et al., 2022)，利用从Timetree (http://www.timetree.org/)查询到大戟科木薯属(6.6 Ma)、杨柳科(43 Ma)、杨柳科刺篱木族(11.9 Ma)的化石时间进行标定。MCMC链长设置为25 000 000代，取样频率为1 000。完成所有参数设置后生成xml文件，使用BEAST v1.8.4运行xml文件。运行结束后得到log文件，在Tracer v1.7中查看Tracer分布图以及有效取样大小(effective sample size, ESS)，若ESS值<200，通过修改MCMC代数使ESS值>200，确定运行参数已收敛。用TreeAnnotator v1.8.4的Maximum clade credibility tree模型运行结果文件，保留Median heights值，运行结束后在Figtree v1.4.3中查看并输出时间树。

2 结果与讨论

2.1 石栗叶绿体基因组特征

通过对石栗叶绿体基因的组装和注释得到完整的叶绿体基因组圈图，表明石栗叶绿体基因组与其它大多数被子植物相同，为典型的四段式结构，包含单拷贝区LSC和SSC以及1对反向重复序列IRA和IRB (图2)。石栗叶绿体基因组总长度为163 298 bp，其中2个IR长度均为26 748 bp，各占石栗叶绿体基因组的16.38%，LSC和SSC的长度分别为91 301和18 501 bp，分别占石栗叶绿体基因组的55.91%和11.33%。石栗叶绿体基因组共编码131个基因，包括8个rRNA基因, 37个tRNA基因, 86个蛋白编码基因(表1)。24个基因(*ndhB*、*rps12*、*trnL*-GAU、*rrn23*、*trnA*-UGC、*ndhA*、*rrn23*、*trnA*-UGC、*trnL*-GAU、*rps12*、*ndhB*、*rpl2*、*trnK*-UUU、*rps16*、*trnG*-UCC、*atpF*、*rpoC-1*、*trnL*-UAA、*trnV*-UAC、*rps12*、*petB*、*petD*、*rpl16*和*rpl2*)有1个内含子，2个基因(*clpP*和*ycf3*)有2个内含子。17个基因含有2个拷贝，包含6个CDS基因(*rpl2*、*rpl23*、*rps7*、*rps12*、*ndhB*和*ycf2*)、7个tRNA基因(*trnL*-CAU、*trnL*-CAA、*trnL*-GAU、*trnV*-GAC、*trnA*-UGC、*trnR*-ACG和*trnN*-GUU)和4个rRNA基因(*rrn16*、*rrn23*、*rrn4.5*和*rrn5*)。*trnK*-UUU与*matK*以及*psbD*与*psbC*之间均有部分重叠，重叠序列长度分别为1 542和53 bp。

石栗叶绿体基因组的LSC和SSC的GC含量分别为33.0%和29.6%，IR的CG含量为42.4%，总体GC含量为35.7%。相较于LSC和SSC，IR的CG含量明显较高，其中8个rRNA分布于IR是造成此现象的重要原因。

2.2 重复序列分析

微卫星DNA又称简单重复序列，是指基因组中由1–6个核苷酸组成的基本单位重复多次构成的一段DNA，广泛分布于基因组的不同位置。通过对石栗叶绿体基因组SSR进行统计，发现十核苷酸以下简单重复序列的重复单元共有145个，形式有单核苷酸、二核苷酸、三核苷酸及四核苷酸，在基因组中数目分别为80、53、10和2 (表2)。单核苷酸多以A和T、二核苷酸多以AT和TA为重复单位，AG和CT作为重复单元

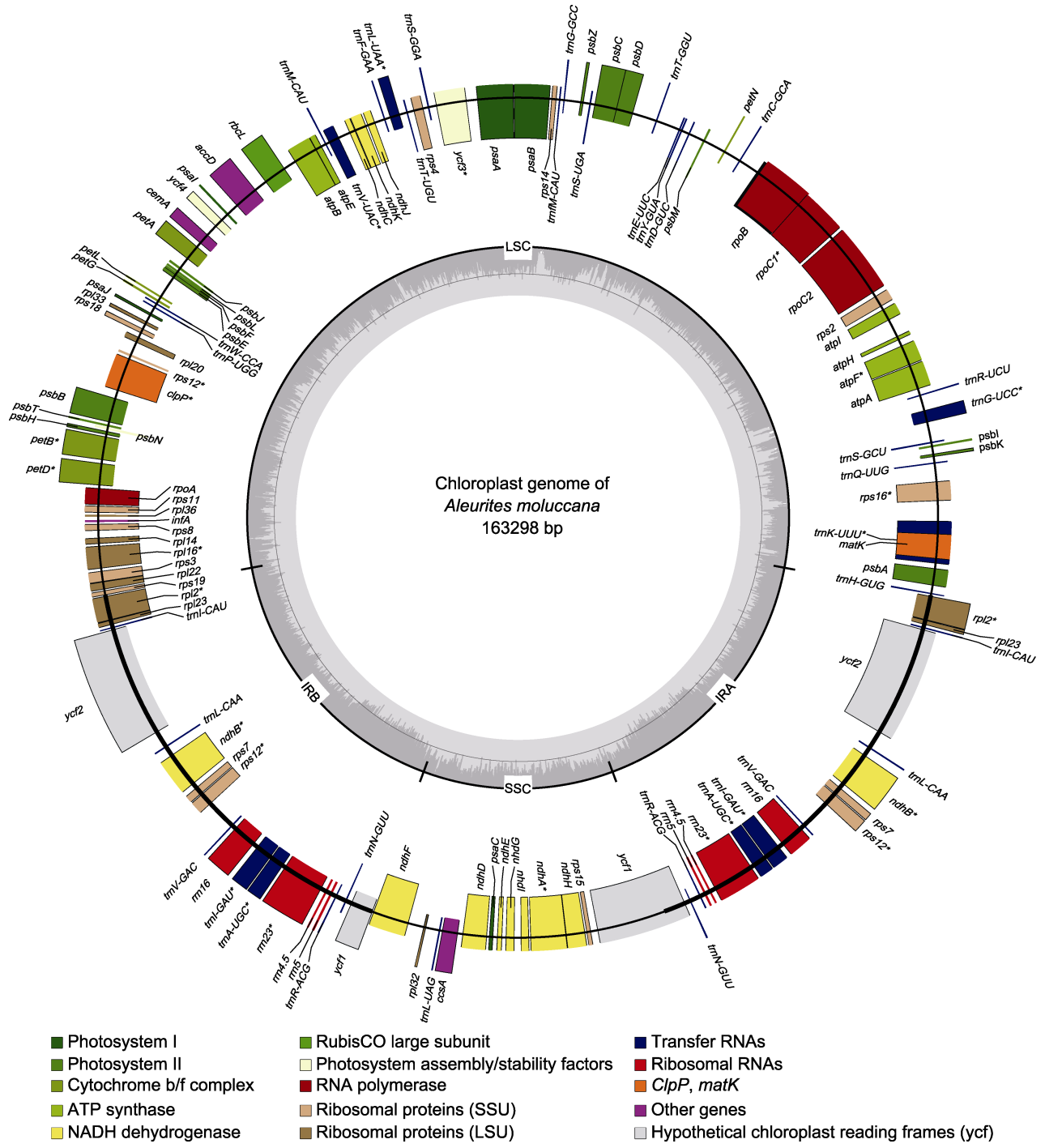

图2 石栗叶绿体基因组图谱

标注在大环外部的基因按照顺时针方向转录，标注在大环内部的基因按照逆时针方向转录。不同颜色代表基因功能不同。内环阴影部分代表石栗叶绿体基因组的GC组成。LSC: 大单拷贝区; SSC: 小单拷贝区; IRA: 反向重复区A; IRB: 反向重复区B。* 含有内含子的基因。

Figure 2 The chloroplast genome map of *Aleurites moluccana*

Genes on the outside of the large circle are transcribed clockwise and those on the inside are transcribed counterclockwise. The genes are color-coded based on their function. The dashed area on the inside represents the GC composition of the *A. moluccana* chloroplast genome. LSC: Large single-copy region; SSC: Small single-copy region; IRA: Inverted repeat region A; IRB: Inverted repeat region B. * for genes containing introns.

表1 石栗叶绿体基因组功能基因注释

Table 1 Annotation of functional genes in the chloroplast genome of *Aleurites moluccana*

Categories of genes	Group of genes	Name of genes
Genes for photosynthesis	Subunits of photosystem I	*psaA*, *psaB*, *psaC*, *psaI*, *psaJ*
	Subunits of photosystem II	*psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbJ*, *psbK*, *psbL*, *psbM*, *psbN*, *psbT*, *psbZ*
	Subunits of ATP synthase	*atpA*, *atpB*, *atpE*, *atpF*, *atpH*, *atpI*
	Subunits of cytochrome	*petA*, *petB*, *petD*, *petG*, *petL*, *petN*
	ATP-dependent protease subunits *P* gene	*clpP*
	Large subunits of Rubisco	*rbcL*
	Subunits of NADH dehydrogenase	*ndhA*, *ndhB*, *ndhC*, *ndhD*, *ndhE*, *ndhF*, *ndhG*, *ndhH*, *ndhI*, *ndhJ*, *ndhK*
Self replication	Small subunit of ribosome	*rps2*, *rps3*, *rps4*, *rps7*, *rps8*, *rps11*, *rps12*, *rps14*, *rps15*, *rps16*, *rps18*, *rps19*
	Large subunit of ribosome	*rpl2*, *rpl14*, *rpl16*, *rpl20*, *rpl2*, *rpl23*, *rpl32*, *rpl33*, *rpl36*, *rpl22*
	DNA dependent RNA polymerase	*rpoA*, *rpoB*, *rpoC1*, *rpoC2*
	Ribosomal RNA genes	*rrn5*, *rrn4.5*, *rrn16*, *rrn23*
	Transfer RNA genes	*trnN*-GUU, *trnR*-ACG, *trnH*-GUG, *trnL*-CAU, *trnA*-UGC, *trnL-GAU*, *trnV*-GAC, *trnL*-CAA, *trnL*-CAU, *trnP*-UGG, *trnW*-CCA, *trnM*-CAU, *trnV*-UAC, *trnF*-GAA, *trnL*-UAA, *trnT*-UGU, *trnS*-GGA, *trnfM*-CAU, *trnG*-GCC, *trnS*-UGA, *trnT*-GGU, *trnE*-UUC, *trnY*-GUA, *trnD*-GUC, *trnC*-GCA, *trnR*-UCU, *trnS*-GCU, *trnQ*-UUG, *trnK*-UUU, *trnL*-CAA, *trnV*-GAC, *trnL*-GAU, *trnR*-UGC, *trnL*-UAG, *trnR*-ACG, *trnN*-GUU
Other genes	Maturase	*matK*
	Envelop membrane protein	*cemA*
	Translation initiation factor IF-1	*infA*
	C-type cytochrome synthesis gene	*ccsA*
Unknown function	Conserved open reading frames	*ycf1*, *ycf2*, *ycf3*, *ycf4*, *ycf15*

表2 石栗叶绿体基因组简单重复序列(SSR)信息

Table 2 Information of simple sequence repeat (SSR) identified in the chloroplast genome of *Aleurites moluccana*

SSR repeat type (number of copies)	SSR repeat sequence	Number of copies												
		5	6	7	8	9	10	11	12	13	14	15	16–30	Total
Mononucleotide (80)	A/T						24	17	7	14		3	3	68
	C/G						7	3		2				12
Dinucleotide (53)	AT/TA/AC/CA/AG/GA		18	7		6	4	6		2			2	45
	CG/GC		4	2		1	1							8
Trinucleotide (10)	CGG/AAG	2	2	1										5
	TGT/TTA	1		2										3
	ATT/AAT	1	1											2
Tetranucleotide (2)	TACA/TGGT	1			1									2

较少。复合SSR有31个, 这些SSR在基因组4个区域中均有分布, 大多数集中在LSC, 其次是SSC和IR, 多位于非编码区。在石栗叶绿体基因组中鉴定出50个重复序列, 其中有34个正向重复, 16个回文重复。正向重复较多(占68%), 重复长度在40–111 bp之间。回文重复较少(占32%), 重复长度在48–111 bp之间。无反向重复和互补重复(图3A, B)。

2.3 密码子偏好性分析

石栗叶绿体基因组共编码20种氨基酸, 各种氨基酸的使用频率在1.46%–10.37%之间, 平均值为5%。亮氨酸使用频率最高, 其次为异亮氨酸、丝氨酸和甘氨酸(图4)。甲硫氨酸和色氨酸只使用2个密码子AUG和UGG。亮氨酸、丝氨酸和精氨酸使用6个密码子。天冬氨酸、谷氨酸、组氨酸、赖氨酸、天冬酰胺、半胱氨酸、谷氨酰胺、酪氨酸和苯丙氨酸均使用2个密码子且具有密码子偏好性, 其氨基酸密码子使用频率较高的分别是UGU、GAU、GAA、UUU、CAU、AAA、AAU、CAA和UAU。丙氨酸、甘氨酸、脯氨酸、苏氨酸和缬氨酸使用4个密码子。异亮氨酸使用3个密码子, 使用频率最高的为AUU (表3)。由此可知, 密码子中A和U的使用频率较高, 且密码子末端碱基多使用A和U, 表明密码子对A和U碱基具有偏好性。

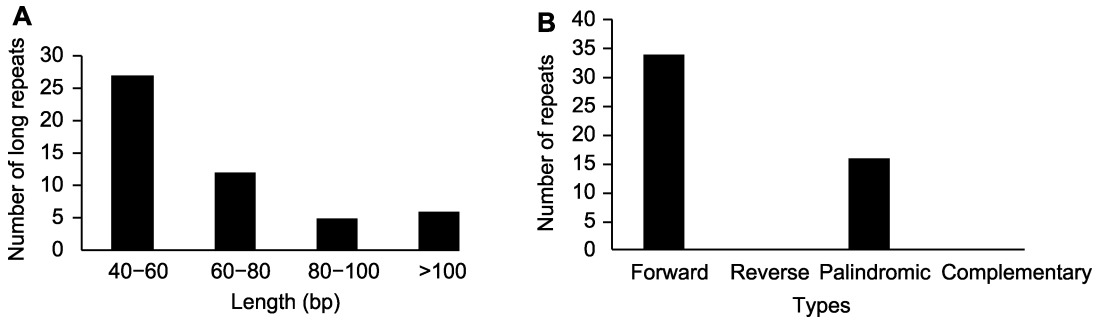

图3 石栗叶绿体基因组重复序列长度及类型
(A) 重复序列长度; **(B)** 重复序列类型

Figure 3 The length and type of repeat sequences in the chloroplast genome of *Aleurites moluccana*
(A) The length of repeat sequences; **(B)** The type of repeat sequences

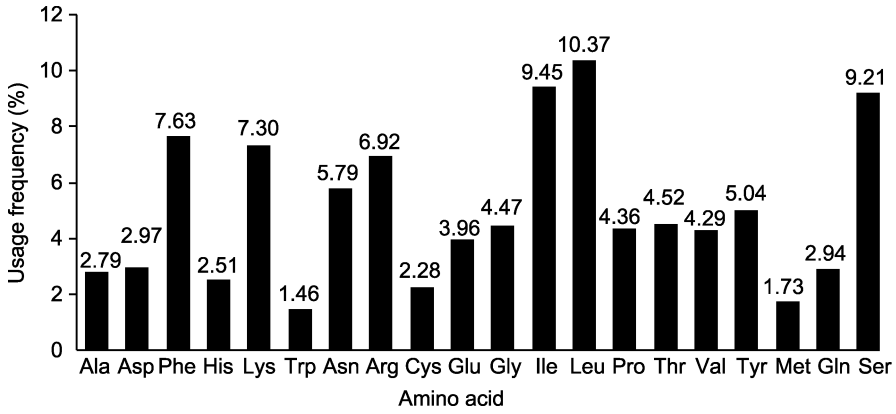

图4 石栗叶绿体基因组不同氨基酸的使用频率

Figure 4 Usage frequency of different amino acid in the chloroplast genome of *Aleurites moluccana*

2.4 共线性分析

共线性(collinearity)是指2个物种具有相同的遗传图谱且它们在相同的位置上有连锁的基因(Liu et al., 2014), 这些基因在结构和功能上具有很高的相似性。近缘物种之间由于分化时间较短, 因而基因发生变异的频率较低, 但也存在基因重排现象。为检测石栗基因组是否发生基因重排及基因缺失、重复、倒位和易位, 我们对石栗及其近缘种油桐、木油桐、东京桐进行共线性分析。结果表明, 除石栗以外其它3个物种基因组LSC和SSC都比较保守, 没有发生明显的基因重排, 相较于其它3个物种, 石栗叶绿体基因组中部分基因发生了明显的重排和倒置(图5, 绿色和紫色部分)。总体而言, 石栗比其它3个物种遗传距离较大, 可能是石栗在漫长的演化过程中发生了遗传突变, 也可能由环境原因导致(Song et al., 2015)。

2.5 IR边界的收缩与扩张

植物叶绿体基因组的IR存在收缩、扩张或缺失的情况, 因此IR区段扩增或减少是叶绿体基因组大小变异的来源。豌豆(*Pisum sativum*)和一些红藻的叶绿体基因组丢失了IR区段, 导致其长度变短。由于反向重复可能有助于稳定叶绿体基因组其它区域, 因此丢失反向重复片段的叶绿体DNA倾向于更多地重排。IR区段主要分布着编码rRNA和tRNA的基因, 其中包括编码16S rRNA、23S rRNA、4.5S rRNA和5S rRNA的基因。通过分析石栗与其近缘种的IR以及IR与两边的单拷贝区临界基因, 发现各区的大小较均衡, 并且在IR长短发生变化的时候, 基因组中的基因发生了改变(图6)。*trnH*基因在石栗、木薯、麻风树和少花橡胶树中位于LSC靠近IRA处, 而油桐和东京桐中则在LSC靠近IRB处。*rpl22*与*trnH*的情况恰恰相反。在石栗和麻风

表3 石栗叶绿体基因组密码子使用偏好性

Table 3 Codon usage bias in the chloroplast genome of *Aleurites moluccana*

Amino acid	Codon	No. of codon	RSCU	Amino acid	Codon	No. of codon	RSCU
Ala	GCU	431	1.20	Cys	UGU	698	1.20
	GCC	306	0.86		UGC	468	0.80
	GCA	427	1.19	Glu	GAA	1416	1.40
	GCG	267	0.75		GAG	611	0.60
Asp	GAU	1073	1.41	Gly	GGU	534	0.93
	GAC	448	0.59		GGC	382	0.67
Phe	UUU	2457	1.26		GGA	803	1.40
	UUC	1452	0.74		GGG	572	1.00
His	CAU	874	1.36	Ile	AUU	1927	1.19
	CAC	412	0.64		AUC	1184	0.73
Lys	AAA	2578	1.38		AUA	1728	1.07
	AAG	1162	0.62	Leu	UUA	1299	1.46
Trp	UGG	748	1.00		UUG	1125	1.26
Asn	AAU	2122	1.43		CUU	1021	1.15
	AAC	843	0.57		CUC	577	0.65
Arg	CGU	341	0.58		CUA	844	0.95
	CGC	246	0.42		CUG	442	0.50
	CGA	590	1.00	Pro	CCU	568	1.02
	CGG	412	0.70		CCC	556	1.00
	AGA	1253	2.12		CCA	699	1.25
	AGG	702	1.19		CCG	408	0.73
Thr	ACU	683	1.18	Met	AUG	887	1.00
	ACC	550	0.95	Gln	CAA	1041	1.38
	ACA	696	1.20		CAG	463	0.62
	ACG	383	0.66	Ser	UCU	1070	1.36
Val	GUU	737	1.34		UCC	840	1.07
	GUC	402	0.73		UCA	1002	1.27
	GUA	667	1.21		UCG	565	0.72
	GUG	393	0.71		AGU	720	0.92
Tyr	UAU	1842	1.43		AGC	521	0.66
	UAC	737	0.57				

RSCU: 相对同义密码子使用度。RSCU: Relative synonymous codon usage.

树中, *rps19*完全位于LSC; 在油桐中, *rps19*则完全位于IRB中; 而在木薯、少花橡胶树和东京桐中, *rps19*位于LSC与IRB的边界上, 分别向IRB延伸了187、97和204 bp。*ycf1*在IRB与SSC和SSC与IRA的边界上均有分布, 在6个物种中, 油桐和东京桐的*ycf1*和*ndhF*基因发生了移位, 在东京桐中, *ycf1*与*ndhF*基因有部分重叠。在IRA与LSC边界上, 只有少花橡胶树的*rpl2*向LSC延伸了114 bp, 其余5个物种中均呈现向IRA收缩的现象。

2.6 系统发育分析

以芸香科的来檬(*C. aurantifolia*)作为外类群, 基于植物叶绿体蛋白编码序列构建贝叶斯和最大似然系统发育树(图7, 图8)。贝叶斯系统发育树的各节点支持率为75%–100%, 显示石栗与油桐属的油桐和木油桐、东京桐以及麻风树聚为一支, 说明石栗属与油桐属和东京桐属的亲缘关系较近。利用最大似然法构建的系统进化树拓扑结构与贝叶斯系统进化树完全一致。

2.7 年龄估计与化石定年

利用BEAST v1.8.4软件宽松分子钟模型计算大戟科的分化时间(图9)。分化时间估算结果表明，大戟科于66 Ma (95% highest posterior density [HPD] interval: 65.8–66.19 Ma)开始出现，石栗属与油桐属、东京桐属的分化时间为25.94 Ma (95% HPD: 24.71–63.32 Ma)，油桐属出现的时间为11.45 Ma (95% HPD: 0.47–21.77 Ma)，油桐属与东京桐属的分化时间为18.83 Ma (95% HPD: 4.61–31.12 Ma)，橡胶属出现的时间为2.42 Ma (95% HPD: 0.43–4.5 Ma)，杨柳科出现的时间为43 Ma (95% HPD: 42.8–43.19 Ma)，蕺菜属与刺篱木属的分化时间为14.7 Ma (95% HPD: 0.4–27.77 Ma)，且蕺菜属开始分化的时间为1.81 Ma (95% HPD: 0.04–3.92 Ma)。

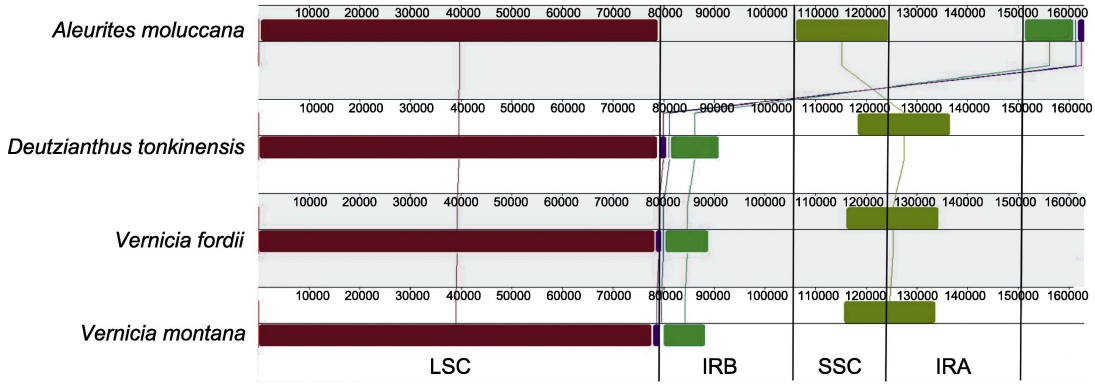

图5 石栗叶绿体基因组共线性分析
LSC、SSC、IRA和IRB同图2。

Figure 5 Collinear analysis of *Aleurites moluccana* chloroplast genome
LSC, SSC, IRA, and IRB are the same as in Figure 2.

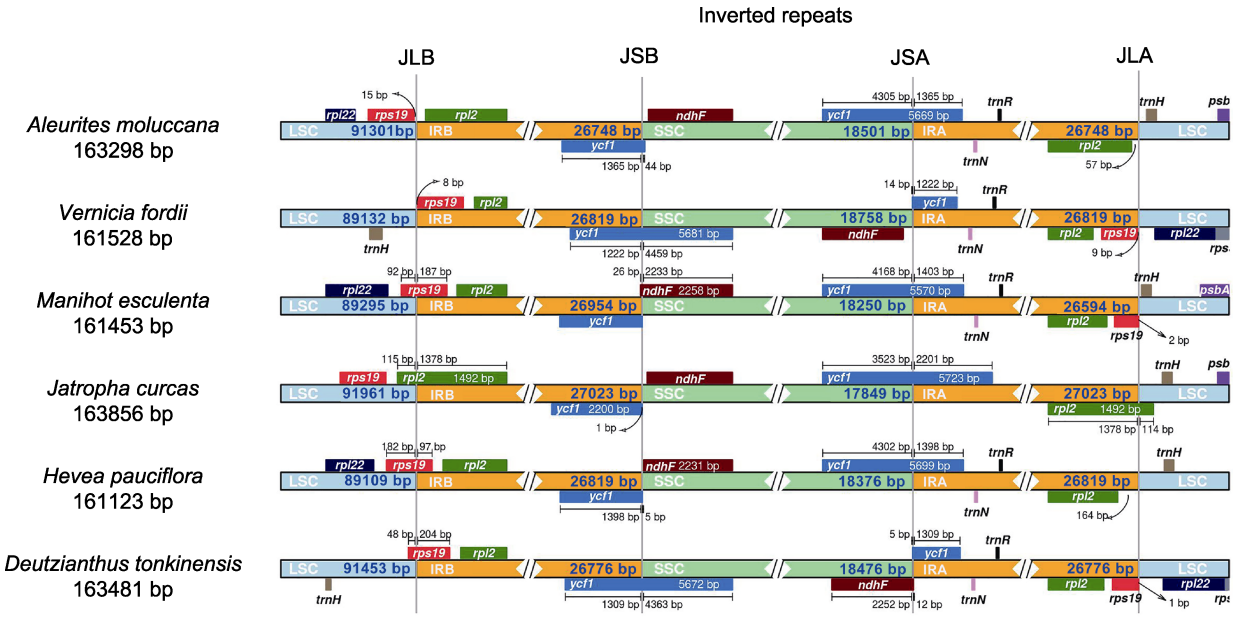

图6 石栗叶绿体基因组反向重复区的收缩与扩张
LSC、SSC、IRA和IRB同图2。JLB: LSC与IRB的边界; JSB: SSC与IRB的边界; JSA: SSC与IRA的边界; JLA: LSC与IRA的边界

Figure 6 Contraction and expansion of inverted repeat region in the chloroplast genome of *Aleurites moluccana*
LSC, SSC, IRA, and IRB are the same as in Figure 2. JLB: Boundary between LSC and IRB; JSB: Boundary between SSC and IRB; JSA: Boundary between SSC and IRA; JLA: Boundary between LSC and IRA

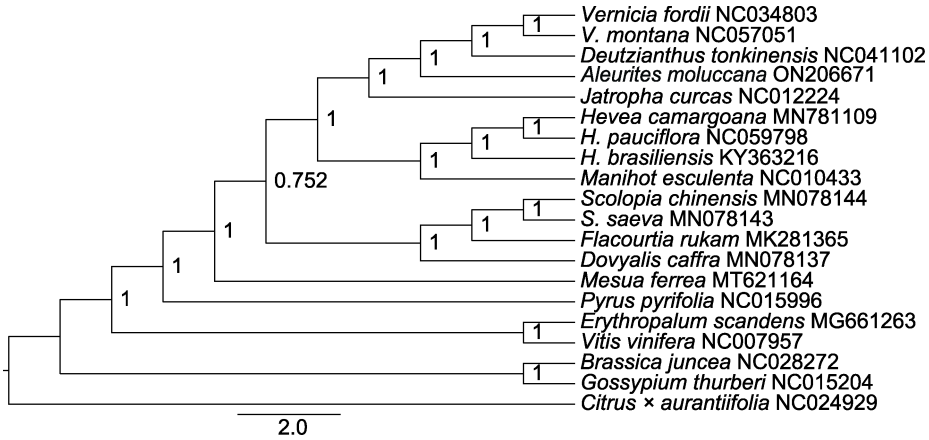

图7 基于蛋白编码序列构建的石栗与其它19个物种贝叶斯系统发育树
分支上的数值为后验概率。

Figure 7 Bayesian phylogenetic tree of *Aleurites moluccana* and other 19 species based on protein coding sequence
The values on the branch are posteriori probability.

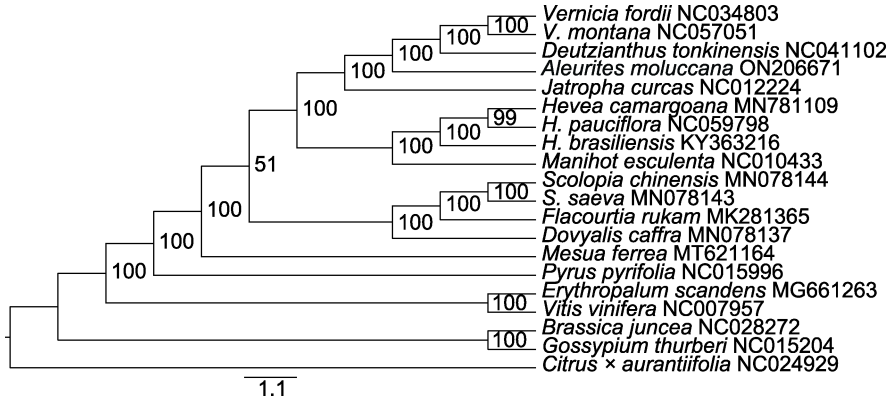

图8 基于蛋白编码序列构建的石栗与其它19个物种最大似然(ML)系统发育树
分支上的数值为后验概率。

Figure 8 Maximum likelihood (ML) phylogenetic tree of *Aleurites moluccana* and other 19 species based on protein coding sequence
The values on the branch are posteriori probability.

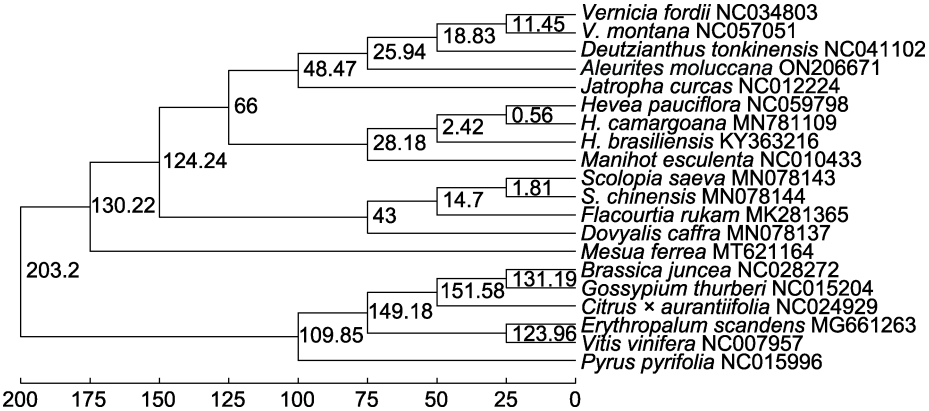

图9 基于宽松分子钟模型构建的石栗与其它19个物种系统发育时间树
分支上的数值为分化时间(单位: 百万年前)。

Figure 9 Phylogenetic dating tree of *Aleurites moluccana* and other 19 species based on relax molecular clock model
The values on the branch are the differentiation time (unit: million years ago).

2.8 讨论

叶绿体基因组为特殊的细胞器基因组，相比核基因组其长度较小，且相对保守。因此叶绿体基因组越来越多地应用于植物系统发育研究(周会等，2014; Li et al., 2017b)。为了解决叶绿体系统发育相关问题，目前已经测定了超过8 400个物种的叶绿体基因组(孙雨晴，2018)，但是对于种类众多的植物群体来说还远远不够。

本研究利用GetOrganelle软件(Jin et al., 2020)基于石栗全基因组数据从头组装获得完整的石栗叶绿体基因组，其与同为大戟科油桐属物种油桐(161 528 bp)、木油桐(164 506 bp)和东京桐属东京桐(163 481 bp)的叶绿体基因组大小差异不大。相比油桐叶绿体基因组，石栗叶绿体基因组SSC及LSC长度较小(Li et al., 2017a)，与橡胶树叶绿体基因组(IR的长度为26 819 bp, LSC为89 281 bp, SSC为18 372 bp)相比，石栗叶绿体基因组的LSC和SSC较长, IR较短(Niu et al., 2020)。研究表明, LSC和SSC在进化过程中会发生扩张和收缩(Martin et al., 2013)，但是由于其叶绿体基因组总体较长，且其它区域并无明显变化，总体来说其LSC长度相对保守，并未发生明显的扩张。石栗叶绿体基因组LSC和SSC的GC含量比IR低，原因是石栗叶绿体基因组编码的rRNA均位于IR。14个基因包含1个内含子, 2个基因包含2个内含子；大戟(*Euphorbia pekinensis*)的5个基因包含9个内含子(Wang et al., 2022)，说明不同物种间内含子位置及数量存在差异。同时还发现石栗与玄参科地黄属叶绿体内含子分布相似，由此推测内含子位置与亲缘关系之间并无关联。个别基因间存在重叠现象，重叠碱基较长的为*trnK*-UUU和*matK*，这些区域有可能是石栗叶绿体基因组中的可变剪接区(Fukuda et al., 1999)。通过对石栗叶绿体重复序列进行统计，发现十核苷酸以下的简单重复序列有145个，单核苷酸以A和T、二核苷酸多以AT和TA为重复单元，这与欧亚种葡萄赤霞珠及桑叶葡萄(*Vitis heyneana* subsp. *ficifolia*)叶绿体基因组以AT/TA作为主要重复单元一致(谢海坤等，2017；杨亚蒙等，2019)。复合SSR在基因组4个区域中均有分布，在LSC中较多，可能是由于该区GC含量较低，也有可能是因其非编码基因数较多所致。与其它被子植物相比，石栗叶绿体基因组中串联重复序列的数量较多，如桑叶葡萄、羊草(*Leymus chinensis*)和马尾松(*Pinus massoniana*)叶绿体基因组中分别含有33、49和26个串联重复序列(杨艳婷，2018；罗群凤等，2018；杨亚蒙等，2019)。但在石栗叶绿体基因组重复序列中只包含正向重复和回文重复，无反向和互补重复，此现象在被子植物叶绿体基因组中较少。

基于石栗与其它19个物种的蛋白编码序列构建最大似然和贝叶斯系统发育树，得到完全相同的拓扑结构。由系统树可知，石栗的分化时间比油桐属和东京桐属早，与油桐、木油桐和东京桐为同一分支并形成姐妹群，与其它大戟科物种亲缘关系较远。杨柳科与同属金尾虎目的大戟科聚为一支，杨柳科的4个物种全部聚为一支，且刺篱木属2个物种形成一个单系群。对大戟科及其它科物种进行化石定年分析，由于石栗属没有化石，选取了大戟科其它物种的化石时间进行校准，估算石栗的分化时间比油桐更早，验证了前人的分类学研究结果，支持将石栗从油桐属中分离出来列为独立的石栗属。

参考文献

蔡金标，丁建祖，陈必勇 (1997). 中国油桐品种、类型的分类. 经济林研究 **15**(4), 47–50.

曹晖，肖艳华，王绍云 (2007). 石栗属和油桐属的化学成分和生物活性. 凯里学院学报 **25**(006), 43–45.

陈琴怡 (2017). 两种五加科植物的叶绿体全基因组研究及其系统发育分析. 硕士论文. 杭州: 浙江大学. pp. 22–34.

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.

梁文汇，李开祥，邓力，曾祥艳，邓福春 (2011). 广西生物柴油原料树种石栗的综合评价. 广西林业科学 **40**, 333–335.

凌建群，张新英，陈耀堂 (1995). 油桐、千年桐和石栗的木材比较解剖. 北京大学学报(自然科学版) **31**, 745–751.

刘昌盛，黄凤洪，李重屹，王明霞，南占东，韩伟，廖李 (2008). 木本油料石栗的初步研究. 中国油料作物学报 **30**, 106–107, 111.

罗群凤，冯源恒，贾婕，陈虎，杨章旗 (2018). 马尾松叶绿体基因组测序及特征分析. 广西林业科学 **47**, 7.

苏梦云，周国璋 (1988). 油桐属与石栗属叶绿体的核酸、蛋白质及超微结构的初步研究. 林业科学研究 **1**, 424–427.

孙雨晴 (2018). 四种葱蒜类蔬菜叶绿体DNA提取优化及比较基因组学研究. 硕士论文. 长春: 吉林农业大学. pp. 36–38.

王劲风, 方嘉兴, 刘兴温, 周国璋, 苏梦云, 成小飞 (1986). 油桐属种分类及其品种类型鉴别方法的探讨. 全国林木遗传育种第五次学术报告会论文汇编. pp.119–121.

王磊 (2013). 石栗种子内含物变化及*accD*基因的克隆研究. 硕士论文. 南宁: 广西大学. pp. 45–68.

谢海坤, 焦健, 樊秀彩, 张颖, 姜建福, 孙海生, 刘崇怀 (2017). 基于高通量测序组装赤霞珠叶绿体基因组及其特征分析. 中国农业科学 **50**, 1655–1665.

杨亚蒙, 焦健, 樊秀彩, 张颖, 姜建福, 李民, 刘崇怀 (2019). 桑叶葡萄叶绿体基因组及其特征分析. 园艺学报 **46**, 635–648.

杨艳婷 (2018). 羊草叶绿体全基因组分析及分子标记开发. 硕士论文. 扬州: 扬州大学. pp. 55–65.

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.

周会, 荆胜利, 李刚, 张磊, 覃瑞, 刘虹 (2014). 叶绿体基因组分析在植物系统发育中的应用. 植物学研究 **3**, 1–9.

Cesca TG, Faqueti LG, Rocha LW, Meira NA, Meyre-Silva C, De Souza MM, Quintão NLM, Silva RML, Filho VC, Bresolin TMB (2012). Antinociceptive, anti-inflammatory and wound healing features in animal models treated with a semisolid herbal medicine based on *Aleurites moluccana* L. Willd. Euforbiaceae standardized leaf extract: semisolid herbal. *J Ethnopharmacol* **143**, 355–362.

Daniell H, Lin CS, Yu M, Chang WJ (2016). Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biol* **17**, 134.

Fukuda Y, Tomita M, Washio T (1999). Comparative study of overlapping genes in the genomes of *Mycoplasma genitalium* and *Mycoplasma pneumoniae*. *Nucleic Acids Res* **27**, 1847–1853.

Jin JJ, Yu WB, Yang JB, Song Y, dePamphilis CW, Yi TS, Li DZ (2020). GetOrganelle: a fast and versatile toolkit for accurate *de novo* assembly of organelle genomes. *Genome Biol* **21**, 241.

Krause K (2008). From chloroplasts to “cryptic” plastids: evolution of plastid genomes in parasitic plants. *Curr Genet* **54**, 111–121.

Li PR, Zhang SJ, Li F, Zhang SF, Zhang H, Wang XW, Sun RF, Bonnema G, Borm TJA (2017a). A phylogenetic analysis of chloroplast genomes elucidates the relationships of the six economically important *Brassica* species comprising the triangle of U. *Front Plant Sci* **8**, 111.

Li Z, Long HX, Zhang L, Liu ZM, Cao HP, Shi MW, Tan XF (2017b). The complete chloroplast genome sequence of tung tree (*Vernicia fordii*): organization and phylogenetic relationships with other angiosperms. *Sci Rep* **7**, 1869.

Liu L, Hao ZZ, Liu YY, Wei XX, Cun YZ, Wang XQ (2014). Phylogeography of *Pinus armandii* and its relatives: heterogeneous contributions of geography and climate changes to the genetic differentiation and diversification of Chinese white pines. *PLoS One* **9**, e85920.

Martin G, Baurens FC, Cardi C, Aury JM, D’Hont A (2013). The complete chloroplast genome of banana (*Musa acuminata*, Zingiberales): insight into plastid monocotyledon evolution. *PLoS One* **8**, e67350.

Nie XJ, Lv SZ, Zhang YX, Du XH, Wang L, Biradar SS, Tan XF, Wan FH, Song WN (2012). Complete chloroplast genome sequence of a major invasive species, crofton weed (*Ageratina adenophora*). *PLoS One* **7**, e36869.

Niu YF, Hu YS, Zheng C, Liu ZY, Liu J (2020). The complete chloroplast genome of *Hevea camargoana*. *Mitochondrial DNA Part B* **5**, 607–608.

Quintão NLM, Pastor MVD, de-Souza Antonialli C, da Silva GF, Rocha LW, Berté TE, de Souza MM, Meyre-Silva C, Lucinda-Silva RM, Bresolin TMB, Filho VC (2019). *Aleurites moluccanus* and its main active constituent, the flavonoid 2″-*O*-rhamnosylswertisin, in experimental model of rheumatoid arthritis. *J Ethnopharmacol* **235**, 248–254.

Radunz A, He P, Schmid GH (1998). Analysis of the seed lipids of *Aleurites montana*. *Z Naturforsch C* **53**, 305–310.

Reback RG, Kapgate DK, Wurdack K, Manchester SR (2022). Fruits of euphorbiaceae from the late cretaceous Deccan intertrappean beds of India. *Int J Plant Sci* **183**, 128–138.

Shaw J, Lickey EB, Schilling EE, Small RL (2007). Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. *Am J Bot* **94**, 275–288.

Song Y, Dong WP, Liu B, Xu C, Yao X, Gao J, Corlett RT (2015). Comparative analysis of complete chloroplast genome sequences of two tropical trees *Machilus yunnanensis* and *Machilus balansae* in the family Lauraceae. *Front Plant Sci* **6**, 662.

Villarante NR, Davila RAE, Sumalapao DEP (2018). Removal of lead (II) by Lumbang, *Aleurites moluccana* activated carbon carboxymethylcellulose composite crosslinked with epichlorohydrin. *Orient J Chem* **34**, 693–703.

Villarante NR, Ibarrientos CH (2021). Physicochemical

characterization of candlenut (*Aleurites moluccana*)-derived biodiesel purified with deed eutectic solvents. *J Oleo Sci* **70**, 113–123.

Wang YL, Jian X, Wang S (2022). Characterization of the complete chloroplast genome of Rupr. (Euphorbiaceae). *Mitochondrial DNA Part B* **7**, 1550–1552.

Zhang QY, Chen X, Guo MB, Guo R, Xu YP, Yang M, Guo HY (2017). Screening and development of chloroplast polymorphic molecular markers on wild hemp (*Cannabis sativa* L.). *Mol Plant Breed* **15**, 979–985.

Analysis of Chloroplast Genomes of *Aleurites moluccana*

Jinbo Bao, Zhijie Ding, Haoyu Miao, Xueli Li, Shuxian Ren, Ruoyan Jiao, Hao Li
Qianqian Deng, Yingzi Li, Xinmin Tian*

Xinjiang Key Laboratory of Biological Resources and Genetic Engineering, College of Life Sciences and Technology, Xinjiang University, Urumqi 830017, China

Abstract *Aleurites moluccana* is an evergreen broad-leaved tree of the genus *Aleurites* in the family Euphorbiaceae, with energy, medicinal and ornamental values. To fill the gap in study of the chloroplast genome of *A. moluccana*, we assembled and annotated the chloroplast genome of *A. moluccana* by next-generation high-throughput whole genome sequencing, and performed genomic characterization and phylogenetic analysis. The results showed that the chloroplast genome of *A. moluccana* exhibited typical quadripartite and circular structures with a total length of 163 298 bp, the length of LSC, SSC, and IR was 91 301, 18 501, and 26 748 bp, respectively. It contains 131 genes, including 8 rRNA genes, 37 tRNA genes and 86 protein coding genes. A total of 145 SSR loci were found, with mononucleotide, dinucleotide, trinucleotide and tetranucleotide repeat units, and the numbers detected were 80, 53, 10, and 2, respectively. The results of collinearity analysis showed that the chloroplast genome of *A. moluccana* has the phenomenon of gene inversion and rearrangement. Phylogenetic trees were constructed using the maximum likelihood and Bayesian methods. It was found that *A. moluccana* was closely related to *Vernicia fordii* and *Deutzianthus tonkinensis*, and formed a sister group. The results of the dating analysis using fossil time showed that the differentiation time of the *Aleurites*, *Vernicia* and *Deutzianthus* was 25.94 Ma (95% HPD: 24.71–63.32 Ma). This study enriched the genomic information of *A. moluccana* and provided basic genetic data for the development and utilization of *A. moluccana* germplasm resources, as well as a reference for species identification and phylogenetic study of the *Aleurites*.

Key words *Aleurites moluccana*, chloroplast genome, phylogeny

Bao JB, Ding ZJ, Miao HY, Li XL, Ren SX, Jiao RY, Li H, Deng QQ, Li YZ, Tian XM (2023). Analysis of chloroplast genomes of *Aleurites moluccana*. *Chin Bull Bot* **58**, 248–260.

* Author for correspondence. E-mail: tianxm06@lzu.edu.cn

(责任编辑：白羽红)