江苏农业学报(*Jiangsu J.of Agr.Sci.*),2023,39(2):504-517
http://jsnyxb.jaas.ac.cn

赵振宁,孙浩田,宋雨茹,等. 山楂属植物叶绿体基因组特征与密码子偏好性分析[J].江苏农业学报,2023,39(2):504-517.
doi:10.3969/j.issn.1000-4440.2023.02.024

山楂属植物叶绿体基因组特征与密码子偏好性分析

赵振宁[1], 孙浩田[2], 宋雨茹[1], 余 潇[3]
(1.西南林业大学林学院,云南 昆明 650224; 2.西南林业大学生态与环境学院,云南 昆明 650224; 3.湖北工程学院建筑学院,湖北 孝感 432000)

摘要: 为明确山楂属植物叶绿体基因组结构与编码蛋白质的基因密码子偏好性特征,本研究利用第二代高通量测序技术对云南山楂[*Crataegus scabrifolia* (Franch.) Rehd.]的叶绿体基因组进行测序、组装和注释,并对山楂属11个种植物的叶绿体基因组结构、遗传多样性以及密码子偏好性进行了分析。结果显示,山楂属植物的叶绿体基因组长度为159 607~159 875 bp,G+C含量为36.6%~36.7%,为标准的四分体结构,G+C含量和结构变异均保守,边界扩张收缩稳定,未发现基因组的倒置和重排现象,11个种植物的简单重复序列和离散重复序列的种类和数量存在一定的差异。综合中性绘图分析、有效密码子数分析(*ENC*-plot)、奇偶校验分析(PR2-plot)和对应性(COA)分析的结果,发现山楂属植物叶绿体基因组密码子使用不但受到碱基突变的影响,还受到选择压力的深刻影响。对叶绿体基因组的最优密码子进行筛选,最优密码子数量为17~20个,其中 *C. kansuensis*、*C. oresbia*、*C. pinnatifida* 的最优密码子数量最多,*C. marshallii* 的最优密码子数量最少,分析它们的最优密码子数据发现,山楂属植物的最优密码子大多以A或U作为第三位碱基。基于CDS(蛋白质编码序列)和叶绿体全基因组构建的系统发育关系既具有一定的相似性,也存在一些差异。本研究结果为山楂属植物的系统发育研究和分子标记开发等工作提供了参考依据。

关键词: 山楂属; 叶绿体基因组; 密码子偏好性; 系统进化
中图分类号: S661.5 **文献标识码:** A **文章编号:** 1000-4440(2023)02-0504-14

Chloroplast genome characteristics and codon usage bias analysis of *Crataegus* L.

ZHAO Zhen-ning[1], SUN Hao-tian[2], SONG Yu-ru[1], YU Xiao[3]
(1.*College of Forestry, Southwest Forestry University, Kunming* 650224, *China*; 2.*College of Ecology and Environment, Southwest Forestry University, Kunming* 650224, *China*; 3.*School of Architecture, Hubei Engineering University, Xiaogan* 432000, *China*)

Abstract: In order to clarify the chloroplast genome structure and codon usage bias of *Crataegus*, this study used the next-generation sequencing to sequence, assemble and annotate the chloroplast genome of *Crataegus scabrifolia* (Franch.) Rehd., and analyzed the chloroplast genome structure, genetic diversity and codon preference of 11 species of *Crataegus*. The results showed that the length of chloroplast genome was between 159 607 bp and 159 875 bp, the G+C content and structural variation were conservative, the G+C content was between 36.6% and 36.7%, the boundary expansion and contraction were stable, no inversion and rearrangement of the genome were found, and there were differences in the type and number of simple sequence repeats and interspersed repeated sequences. Based on the results of neutrality plot analysis, *ENC*-plot, PR2-plot and correspondence analysis, it was found that the chloroplast genome codon us-

收稿日期: 2022-10-17
基金项目: 云南省第二次国家重点保护野生植物资源调查项目(09930-216304);2021年度云南省大学生创新创业国家级项目(202110677046)
作者简介: 赵振宁(2003-),男,山东泰安人,本科,主要从事植物生物信息学研究。(E-mail)zzn1529370396@163.com
通讯作者: 余 潇,(E-mail)yuxiao19920215@163.com

age in *Crataegus* was not only affected by base mutation, but also by selective pressure. The optimal codons of the chloroplast genome were screened, and the optimal number of codons was between 17 and 20. *C. kansuensis*, *C. oresbia*, and *C. pinnatifida* had the largest number of optimal codons, and *C. marshallii* had the least number of optimal codons. The analysis of their optimal codon data revealed that the optimal codons of *Crataegus* mostly used A or U as the third base. The phylogenetic relationships constructed based on protein coding sequence and complete chloroplast genome had certain similarities and differences. The results of this study can provide a reference for the phylogenetic research and molecular marker development of *Crataegus*.

Key words: *Crataegus* L.; chloroplast genome; codon usage bias; system evolution

山楂属(*Crataegus* L.)为蔷薇科中起源相对古老的属,多为小乔木或落叶灌木,主要分布于温带地区。山楂属植物有着非常高的经济价值,研究结果表明,山楂作为果树在中国的种植历史可追溯至汉代[1]。山楂的果实含有丰富的营养物质,具有健胃消食、抗菌消炎等功效,是一种优良的水果[2]。除了作为经济果树,山楂还是一类出色的园林景观植物和街道绿化树种。通常认为,山楂属中有 18 个种原产于中国,山楂属植物中广泛存在的无融合生殖和种间杂交现象使其外形特征发生了高度变异[3],进而为山楂属植物的传统分类学鉴定造成困难。

叶绿体是植物细胞中重要的细胞器之一,对于研究植物体的光合作用和生长发育具有非常重要的意义。叶绿体基因组是独立于核基因组的母系遗传,其核苷酸置换率与核基因组及线粒体基因组相比更适宜应用于多层次的系统发育研究[4]。随着第二代高通量测序技术的不断完善,针对叶绿体基因组的报道也逐渐增多,目前的研究结果表明,陆地高等植物的叶绿体基因组长度一般介于 120~200 kb,包含大单拷贝区(LSC)、小单拷贝区(SSC)、反向重复区 a(IRa)和反向重复区 b(IRb)。密码子偏好性是指编码相同氨基酸的同义密码子频率存在差异[5],这种现象普遍出现在所有原核生物和真核生物中[6]。一般来说,密码子使用模式能够反映基因组的起源和进化模式,不同的基因组有其独特的密码子使用偏好性,这也使得解释这种偏好性目前还存在一定的困难[7-8]。

山楂属植物具有出色的经济价值和科研价值,目前已有许多针对山楂属植物的相关研究。例如,有许多学者围绕山楂属植物的营养价值进行了相关研究,均发现其有着良好的营养价值和抗氧化活性[9-12],在分子层面,张枭等[13]利用 SSR 分子标记构建了部分山楂属植物的分子条形码,为山楂属植物的资源鉴定提供了分子层面的手段,Liston 等[3]基于叶绿体基因组和 257 个核基因组对山楂属植物亚属间的杂交状况进行了评估,证实了杂交在山楂进化中的重要作用。具体到叶绿体基因组层面,近年来,针对山楂属植物叶绿体基因组的研究正逐渐被重视,部分山楂属植物的叶绿体基因组数据相继被发表在国家生物技术信息中心(National Center for Biotechnology Information, NCBI)公共数据库中,也有学者对其叶绿体基因组进行了属内的比较分析[14-15]。然而,目前针对山楂属植物叶绿体基因组特征和密码子偏好性的综合分析相对较少。本研究拟通过对云南山楂叶绿体基因组的测序、组装和注释,综合分析山楂属 11 个种的植物叶绿体基因组特征、密码子偏好性、最优密码子和系统发育关系,深入研究山楂属植物的叶绿体基因组特征,弥补目前对于山楂属植物密码子特征和偏好性研究的空白。本研究旨在为山楂属植物的叶绿体基因组特征、系统发育关系和密码子偏好性研究提供新的参考依据,以期为山楂属植物的育种和分子标记研究提供参考。

1 材料与方法

1.1 试验材料

本研究所使用的新鲜植物叶片采集于云南省大理白族自治州洱源县罗平山(99°52′19.15″E, 25°59′53.34″N,海拔2 105 m),经西南林业大学标本馆树木学教研室李双智副教授鉴定为蔷薇科山楂属植物云南山楂[*Crataegus scabrifolia* (Franch.) Rehd.]。使用改良过的 CTAB(十六烷基三甲基溴化铵)法[16]从使用硅胶干燥的叶片中提取 DNA,提取后的 DNA 送至天津诺禾致源生物科技有限公司进行叶绿体基因组测序,使用 GetOrganelle 软件[17]组装得到完整的叶绿体基因组,并使用拼接路径可视化软件 Bandage[18]验证其成环性。以山楂[*Crataegus pinnatifida*(NC_065486)]叶绿体基因组为参考,使

用 CPGAVAS2 在线工具(http://www.herbalgenomics.org/cpgavas/)[19]对云南山楂叶绿体基因组进行注释,并使用 Geneious Prime 软件[20]对其进行手动调整。注释过的云南山楂叶绿体基因组上传到 GenBank 公共数据库,登录号为 OP021659,其余 10 个山楂属植物叶绿体基因组下载于 NCBI 公共数据库(https://www.ncbi.nlm.nih.gov/)(表 1)。

表 1 山楂属植物叶绿体基因组信息

Table 1 Complete chloroplast genome sample information of *Crataegus*

编号	物种	登录号	长度(bp)
1	*Crataegus maximowiczii*	NC_065485	159 875
2	*Crataegus kansuensis*	NC_039374	159 865
3	*Crataegus oresbia*	NC_065671	159 851
4	*Crataegus chungtienensis*	NC_065670	159 847
5	*Crataegus rhipidophylla*	NC_062345	159 786
6	*Crataegus hupehensis*	NC_054155	159 766
7	*Crataegus cuneata*	NC_058896	159 730
8	*Crataegus marshallii*	MK920293	159 660
9	*Crataegus pinnatifida*	NC_065486	159 656
10	*Crataegus scabrifolia*	OP021659	159 637
11	*Crataegus bretschneideri*	MW963339	159 607

1.2 试验方法

1.2.1 重复序列分析 简单重复序列(Simple sequence repeat,SSR)在植物叶绿体基因组中有着广泛分布,其作为一种重要的分子标记常被用作鉴定植物品种和构建 DNA 指纹图谱[21]。使用 MISA-web(http://webblast.ipk-gatersleben.de/misa/)对山楂属植物简单重复序列的种类和数量进行在线分析[22],将单核苷酸、二核苷酸、三核苷酸、四核苷酸、五核苷酸、六核苷酸参数分别设置为 10、5、4、3、3、3,相邻 SSR 间的最小距离为 100 bp。

使用 REPuter 在线工具(https://bibiserv.cebitec.uni-bielefeld.de/reputer)分别鉴定 11 种山楂属植物的离散重复序列[23],设置参数:海明距离(Hamming distance)为 3,鉴定类型选择正向重复序列(Forward repeat, F)、回文重复序列(Palindromic repeat, P)、反向重复序列(Reverse repeat, R)和互补重复序列(Complement repeat, C)4 种,最小重复长度 30 bp,最大重复长度 300 bp。

1.2.2 边界扩张收缩分析 叶绿体基因组为环状结构,分为 4 个区域,分别为大单拷贝区(LSC)、小单拷贝区(SSC)、反向重复区 a(IRa)和反向重复区 b(IRb),其中反向重复区相对比较保守,其收缩与扩张会影响叶绿体基因组 G+C 含量和基因组大小,边界扩展和收缩能够展现植物的遗传进化[24];分析叶绿体基因组区域边界的信息,对揭示叶绿体基因组的结构差异和进化关系具有重要的参考价值[25]。使用在线工具 CPJSdraw(http://cloud.genepioneer.com:9929)对注释过的山楂属植物叶绿体基因组边界可视化,分析其边界的扩张收缩情况。

1.2.3 共线性比较分析 以山楂属 11 个种的植物叶绿体基因组为研究对象,利用 MAUVE(http://darlinglab.org/mauve/mauve.html)工具对多重基因组的保守区域、局部共线性和基因组重排倒置现象进行鉴定,用以阐述山楂属植物的叶绿体在物种演化过程中发生的结构变异事件[26]。

1.3 密码子偏好性分析

1.3.1 密码子相关参数的计算 使用 Geneious Prime 软件手动提取每个山楂属植物叶绿体基因组中的蛋白质编码序列(Coding sequence, CDS),由于编码长度较短的蛋白质的基因会使密码子偏好性的数据存在较大的估计误差,因此在统计密码子偏好性时,常去除长度在 300 bp 以下的序列,从而避免产生统计误差[27],本研究筛选了山楂属植物叶绿体基因组中具有代表性的 48 个 CDS。利用 CUSP 在线工具(http://www.Bioinformatics.nl/emboss-explorer/)和 Condon W 1.4.2 统计得到了叶绿体基因组的相对同义密码子使用度(*RSCU*),密码子第一、第二和第三位的 G+C 含量(GC_1、GC_2、GC_3)等一系列信息。

1.3.2 中性绘图分析 使用 GC_1 与 GC_2 的平均值(GC_{12})与 GC_3 作为数据绘制中性对比图,中性对比图可以用来检测密码子突变压力和选择压力的平衡,从而揭示 GC_{12} 和 GC_3 的关系[28]。在密码子偏好中性对比中,每个离散点表示 1 个基因,若 GC_{12} 与 GC_3 为中性,则这些点应位于对角线上,若不为中性,这些点应出现在横坐标的平行线上[29]。

1.3.3 *ENC*-plot 分析 有效密码子数分析(*ENC*-plot)用于分析密码子使用受到选择压力和突变压力的影响程度,根据各组基因密码子的 GC_3 和有效密码子数(*ENC*),首先计算出预期 *ENC*(预期*ENC*=

GC_3+2+29/[GC_3^2+(1−GC_3)^2]),然后使用 R 语言绘制 *ENC*-plot 图,通过比较预期 *ENC* 与实际 *ENC* 得出突变压力和选择压力对密码子使用偏好性的影响程度[8]。

1.3.4 PR2-plot 分析 奇偶校验分析(PR2-plot)用于展现突变压力与选择压力对于密码子使用的影响程度,分析密码子第三位碱基的 A、T、C、G 含量(分别为 A_3、T_3、C_3、G_3),并分别以 $G_3/(G_3+C_3)$ 和 $A_3/(A_3+T_3)$ 为横坐标和纵坐标进行 PR2-plot 绘图,各个基因的密码子偏好性通过其与中心点的方向和矢量偏差表示,而图中中心点表示 $A=T$ 和 $C=G$,即此时基因的密码子使用无偏好性[30]。

1.3.5 最优密码子确定 最优密码子表示基因组中使用频率最高的密码子,以 *ENC* 为首选标准,将 48 条叶绿体基因组按照 *ENC* 进行排序,*ENC* 最高的 5 个基因组归为高表达基因组,*ENC* 最低的 5 个基因组为低表达基因组。将同时满足高频[*RSCU*(同义密码子相对使用度)>1]和高表达[*ΔRSCU*(同义密码子相对使用度之差)≥0.08]的密码子作为最优密码子。

1.3.6 对应性分析 使用 CodonW 1.4.2 基于 *RSCU* 对山楂属 11 个种进行对应性分析,将山楂属这 11 个种所共有的 48 个编码蛋白质的基因组按照基因功能分为 5 种类型,通过分析其变异情况得到影响其密码子偏好性的主要影响因素。

1.4 系统发育分析

基于山楂属 11 个种构建叶绿体全基因组系统发育树和 CDS 系统发育树。先将山楂属 11 个种植物叶绿体全基因组和 CDS 通过 MAFFT v.7 软件进行比对[31],比对结果通过 trimAl[32]进行修饰,修改后的比对文件基于 RAxMLv.8 中的 GTR+I+G 模型,采用最大似然法进行系统发育分析[33],设置 1 000 次自展值重复。

2 结果与分析

2.1 叶绿体基因组结构

山楂属植物叶绿体基因组呈现标准的四分体结构,分别为大单拷贝区、小单拷贝区、反向重复区 a 和反向重复区 b,叶绿体基因组全长为159 607~159 875 bp(图 1)。LSC 长度为87 601~87 874 bp,SSC 长度为19 139~19 312 bp,单个反向重复区长度为26 347~26 385 bp。各个种的 G+C 含量为 36.6%~36.7%,基因总数为127~132 个,其中 rRNA 数量均为 8 个,tRNA 数量除 *C. scabrifolia* 为 36 个外其余均为 37 个,编码蛋白质的基因数量为83~85 个(表 2)。综合来看,山楂属植物的叶绿体基因组 G+C 含量相近,基因种类和数量相近,未发现 IR 区丢失现象,叶绿体基因组长度变异较小,结构未发现明显差异。

表 2 山楂属植物叶绿体基因组结构信息

Table 2 Chloroplast genome structure information of *Crataegus* species

物种	G+C 含量(%)	tRNA	编码蛋白质的基因数量(个)	基因总数(个)	LSC 长度(bp)	SSC 长度(bp)	IR 长度(bp)
C. maximowiczii	36.6	37	85	132	87 874	19 233	26 384
C. kansuensis	36.6	37	85	132	87 815	19 282	26 384
C. oresbia	36.6	37	84	132	87 819	19 264	26 384
C. chungtienensis	36.6	37	84	132	87 815	19 264	26 384
C. rhipidophylla	36.7	37	83	128	87 777	19 241	26 384
C. hupehensis	36.6	37	85	132	87 852	19 144	26 385
C. cuneata	36.6	37	84	129	87 778	19 184	26 384
C. marshallii	36.6	37	85	132	87 698	19 232	26 365
C. pinnatifida	36.7	37	85	132	87 749	19 139	26 384
C. scabrifolia	36.7	36	83	127	87 730	19 139	26 384
C. bretschneideri	36.6	37	85	131	87 601	19 312	26 347

LSC:大单拷贝区;SSC:小单拷贝区;IR:反向重复区。

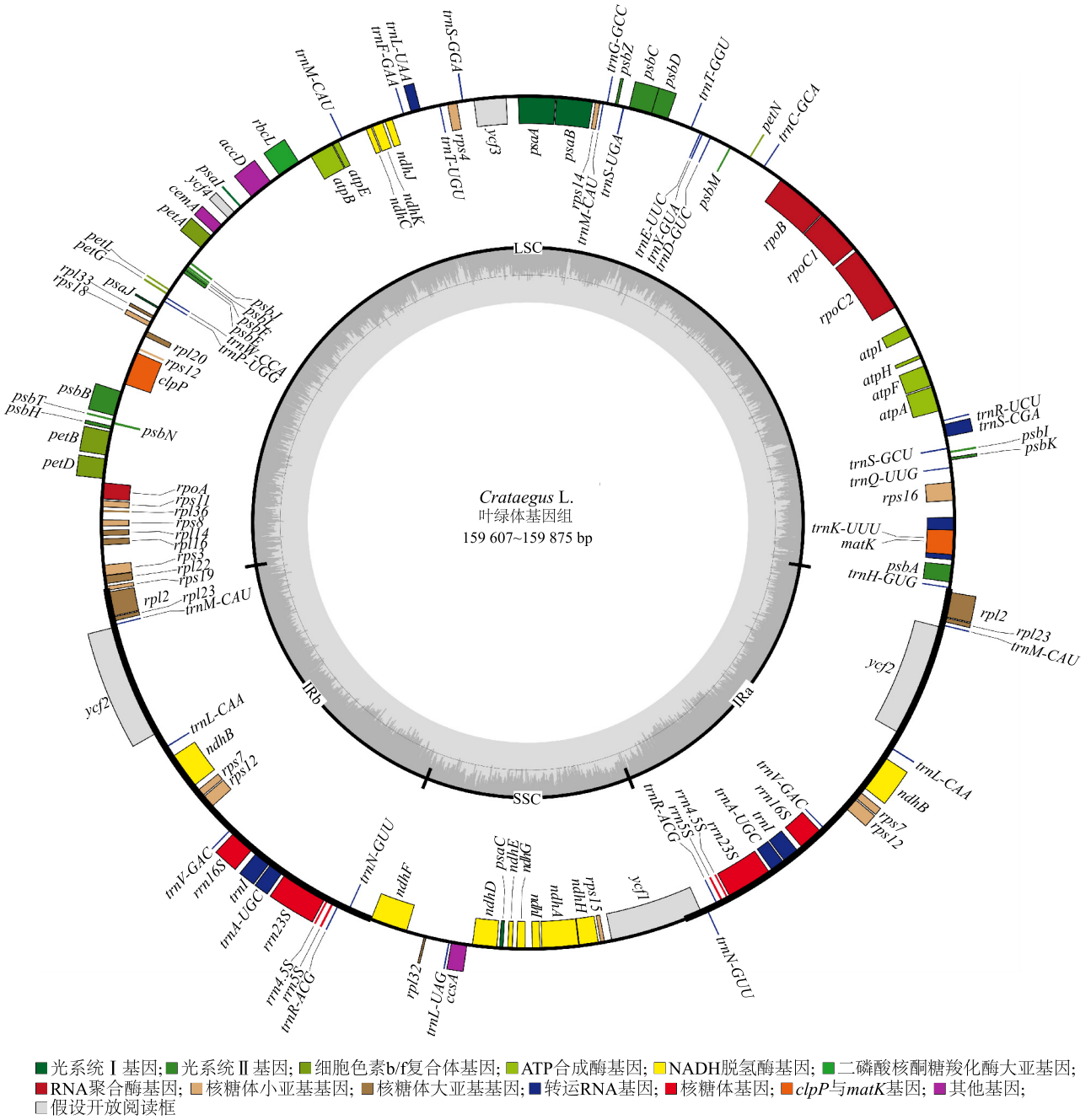


图 1 山楂属植物叶绿体基因组图谱

Fig.1 Chloroplast genome map of *Crataegus*

2.1.1 重复序列分析 如图 2A 所示,在本研究中,单核苷酸、二核苷酸、四核苷酸和复合重复序列均在山楂属植物中被检测到,在本研究所选取的山楂属植物中,检测到的单核苷酸重复序列数量介于41~55,在各个种中单核苷酸重复序列数量均排第一位,而单核苷酸重复序列数量最多的物种为 *C. hupehensis*,最少的物种为 *C. marshallii*。山楂属植物中二核苷酸重复序列数量总体差异不大,*C. oresbia* 被检测到的二核苷酸重复序列数量最少,为 13 个,*C. maximowiczii*、*C. kansuensis*、*C. cuneata*、和 *C. bretschneideri* 数量最多,为 15 个,其余物种则为 14 个。三核苷酸重复序列仅在 *C. hupehensis*、*C. cuneata*、*C. marshallii*、*C. pinnatifida* 和 *C. scabrifolia* 中被检测到,四核苷酸重复序列数量为3~5 个,各物种之间差异不大。五核苷酸重复序列仅在 *C. marshallii* 中被检测到,六核苷酸重复序列仅在 *C. cuneata* 和 *C. marshallii* 中被检测到。这一结

果说明山楂属植物的简单重复序列的类型和数量有部分相似之处,但总体来看也有一定的差异。

使用 REputer 在线工具对 11 种山楂属植物叶绿体基因组的离散重复序列进行分析,统计结果如图 2B 所示,结果显示山楂属植物离散重复序列具有一定相似性,回文重复序列为23~28,正向重复序列为20~29,其中 *C. kansuensis* 的回文重复序列与反向重复序列的数量均为最多,而 *C. marshallii* 的 2 种重复序列的数量均为最少。反向重复序列为3~11 个,其中 *C. kansuensis* 的反向重复序列数量远高于其他 10 个种,为 11 个。互补重复序列在 *C. hupehensis* 中检测到的数量最多,为 5 个,而在 *C. maximowiczii* 与 *C. bretschneideri* 中并未检测出互补重复序列。总的来说山楂属植物的离散重复序列存在着一定的差别。

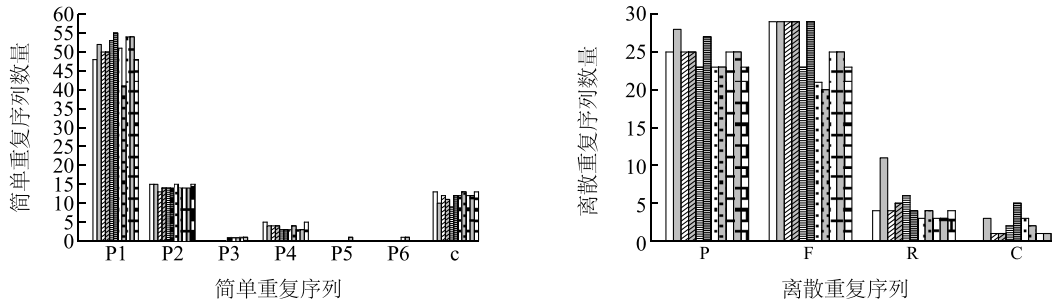


图 A:简单重复序列;图 B:离散重复序列。图 A 中,P1、P2、P3、P4、P5、P6 和 c 分别表示单核苷酸、二核苷酸、三核苷酸、四核苷酸、五核苷酸、六核苷酸和复合重复序列;图 B 中,P:回文重复序列;F:正向重复序列;R:反向重复序列;C:互补重复序列。

图 2 山楂属植物叶绿体基因组重复序列

Fig.2 Repeated sequence of *Crataegus* species chloroplast genome

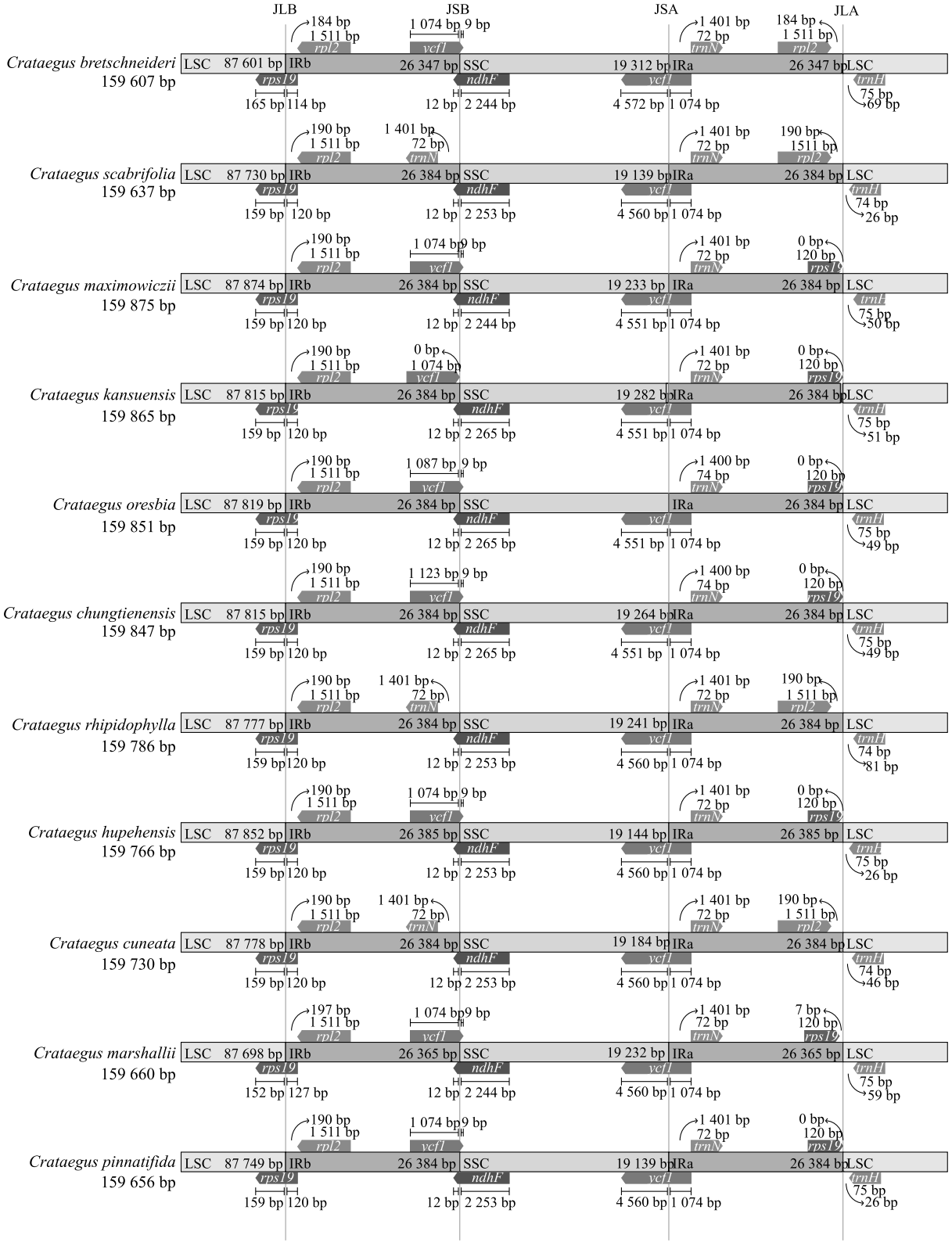
2.1.2 *边界扩张收缩分析* 对山楂属植物的边界扩张收缩分析结果(图 3)表明,山楂属 11 个种植物的大单拷贝区与反向重复区 b 的边界(JLB)均位于 *rps19* 基因中,除 *C. marshallii* 和 *C. bretschneideri* 外,其余 9 个种的 *rps19* 基因均有 120 bp 位于 IRb 区域中;反向重复区 b 与小单拷贝区的边界(JSB)均位于 *ndhF* 中,且 *ndhF* 位于 IRb 的长度均为 12 bp;JSA 均存在于 *ycf1* 基因中,且均有1 074 bp 位于 IRa 中,*rpl2* 为 11 个种植物的共有基因,均位于大单拷贝区与反向重复区 a(JLA)的左侧,其中有 9 个种植物 *rpl2* 基因与 JLA 距离为 190 bp,而 *C. marshallii* 和 *C. bretschneideri* 的 *rpl2* 基因与 JLA 的距离则发生了变异,与其余 9 个种植物略有不同。总的来说,山楂属植物的叶绿体基因组进化关系保守,结构差异较小,边界扩张收缩幅度较为稳定,只发生了较小的变异。

2.1.3 *共线性分析* 使用 Mauve 软件,采用多重基因组比较法对山楂属 11 个种植物的叶绿体基因组进行共线性分析,山楂属植物叶绿体基因组结构与各个基因的排列顺序基本一致,共线性良好,未发现倒置和重排现象,叶绿体基因组之间具有高度相似性。

2.2 密码子偏好性

2.2.1 *密码子组成分析* 在研究密码子的使用偏好性时,*ENC* 常用于评价物种密码子偏好性的大小,其值为20~61,*ENC* 值越大表示密码子的偏好性越弱。一般认为,*ENC* 值在 35 以下时可表明其密码子偏性现象较为显著[34]。由表 3 可知,山楂属 11 个种植物的叶绿体基因组平均 *ENC* 为 46.61~47.55,均大于 35,密码子偏好性较弱,密码子总 G+C 含量与第一、第二、第三位的 G+C 含量均小于 50%,且呈现出$GC_1>GC_2>GC_3$的趋势,说明山楂属植物的叶绿体基因组富含 A 和 T 2 种碱基,且偏好于使用 A、T 作为密码子第三位结尾碱基。

2.2.2 *PR2-plot 绘图分析* 若密码子的偏好性只受突变压力的影响,则 A、T 与 C、G 的使用频率应该是完全相等的。由图 4 可知,图中坐标点的分布并不均匀,可以明显看出,右侧的坐标点多于左侧,下方的坐标点多于上方,而分布于右下角区域的基因数量最多,说明山楂属植物叶绿体基因组密码子第三位碱基对于 T 的使用率大于 A,对于 G 的使用率大于 C,说明其密码子偏好性不只受到突变的影响,而是选择压力和突变压力共同作用的结果。



JLB 表示 LSC 与 IRb 的边界，JSB 表示 SSC 与 IRb 的边界，JSA 表示 SSC 与 IRa 的边界，JLA 表示 LSC 与 IRa 的边界。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(IRb)。

图 3 山楂属植物叶绿体基因组边界

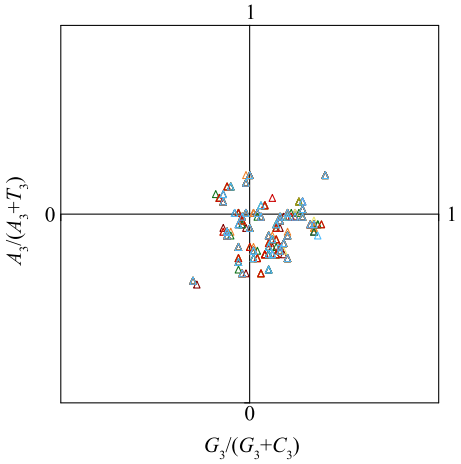
Fig.3 The chloroplast genome boundary analysis of *Crataegus* species

表 3 山楂属植物叶绿体基因组密码子参数特征

Table 3 Condon features of chloroplast genomes of *Crataegus* species

物种	GC_1(%)	GC_2(%)	GC_3(%)	GC_{all}(%)	*ENC*(%)	GC_{3s}(%)
C. maximowiczii	45.82	37.69	29.11	37.55	46.61	29.22
C. kansuensis	45.84	37.73	29.13	37.64	47.32	29.21
C. oresbia	45.86	37.76	29.12	37.56	47.55	29.23
C. chungtienensis	45.88	37.79	29.15	37.58	47.30	29.22
C. rhipidophylla	45.85	37.78	29.16	37.55	46.95	29.25
C. hupehensis	45.88	37.76	29.19	37.65	47.37	29.30
C. cuneata	45.84	37.75	29.18	37.66	47.12	29.33
C. marshallii	45.81	37.74	29.17	37.64	47.28	29.25
C. pinnatifida	45.83	37.75	29.15	37.62	47.43	29.26
C. scabrifolia	45.86	37.86	29.16	37.61	46.75	29.24
C. bretschneideri	45.87	37.77	29.12	37.59	47.32	29.28

GC_1、GC_2、GC_3分别表示密码子第一、第二、第三位碱基的 G+C 含量;GC_{all}表示密码子总的 G+C 含量;*ENC* 表示有效密码子数;GC_{3s}表示同义密码子第三位的 G+C 含量。



△ *C.maximowiczii*; △ *C.kansuensis*; △ *C.oresbia*; △ *C.chungtienensis*; △ *C.rhipidophylla*; △ *C.hupehensis*; △ *C.cuneata*; △ *C.marshallii*; △ *C.pinnatifida*; △ *C.scabrifolia*; △ *C.bretschneideri*

横坐标表示 G、C 碱基偏好性,纵坐标表示 A、T 碱基偏好性。

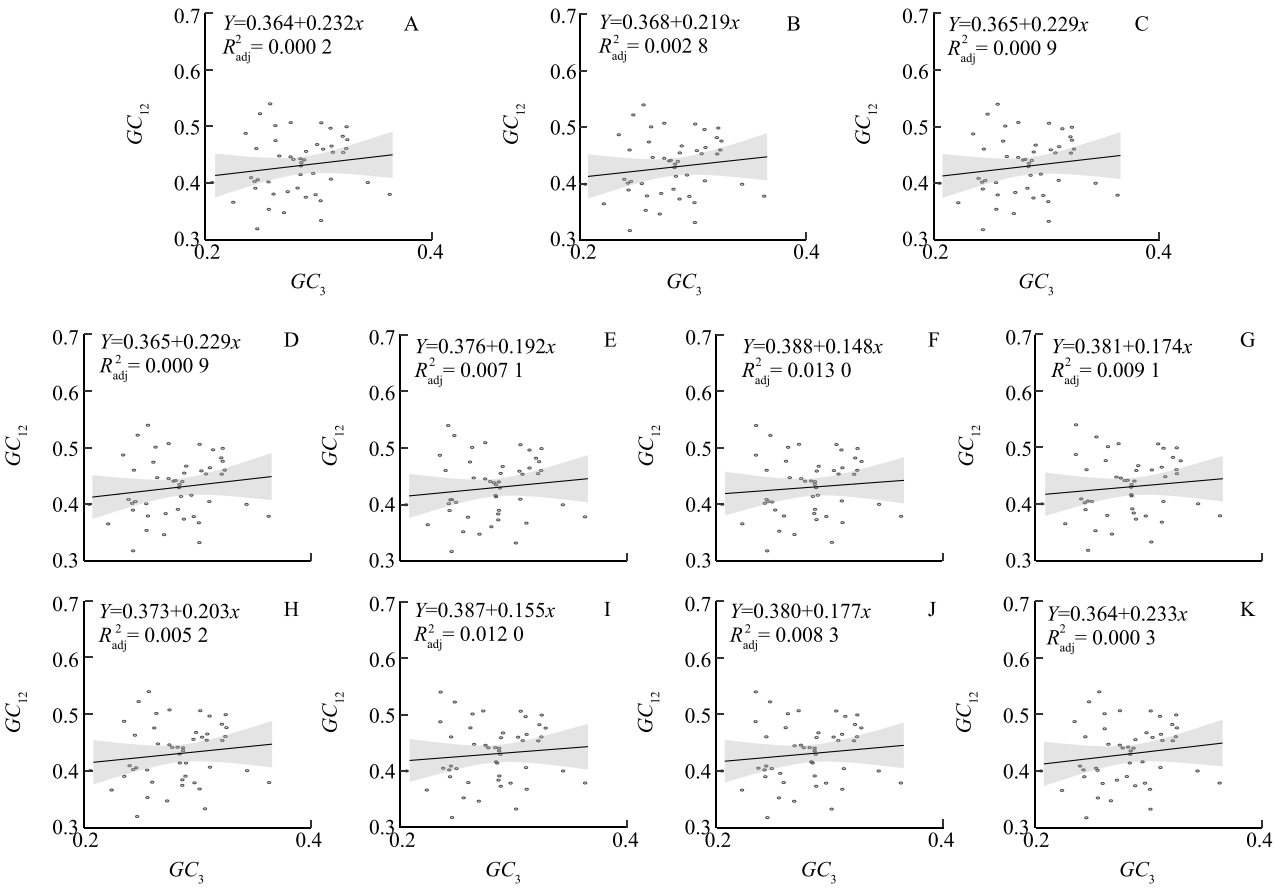
图 4 山楂属植物叶绿体基因组奇偶偏好性

Fig.4 Parity rule 2 analysis of chloroplast genomes from *Crataegus* species

2.2.3 中性绘图分析 山楂属植物中性绘图分析见图 5,各基因的 GC_3取值为20.74%~36.54%,GC_{12}的值则介于31.75%~53.96%,回归系数为0.364~0.388,GC_{12}与 GC_3的相关系数为0.324~0.525,双尾检验均未达到显著水平($P>0.05$),GC_{12}与 GC_3之间相关性不显著,选择压力对其密码子的偏好性具有显著影响,说明山楂属植物叶绿体基因组密码子的第一、第二位碱基与第三位碱基的组成相关性较弱,密码子受选择压力的影响较大。

2.2.4 *ENC*-plot 绘图分析 ENC-plot 绘图能够揭示基因组密码子的 *ENC* 与 GC_3之间的联系,如图 6 所示,坐标点大多分布在标准 *ENC* 曲线下方,且大多与预期 *ENC* 差距很大,即大部分基因的实际 *ENC* 小于预期值,这部分基因主要受到自然选择的影响。仅有少数基因靠近标准曲线,即只有少数基因的密码子偏好性主要受到突变压力的影响。总的来说,在本研究中,自然选择压力是供试山楂属植物叶绿体基因组密码子偏好性的主要影响因素。

2.2.5 山楂属植物最优密码子 对 48 个 CDS 基因按照 *ENC* 进行排序,根据高表达基因和低表达基因中密码子的 *RSCU* 和△*RSCU* 来确定其最优密码子,筛选得到的最优密码子如表 4 所示,最优密码子数量介于 17~20 个,*C. kansuensis*、*C. oresbia*、*C. pinnatifida* 的最优密码子数量最多,*C. marshallii* 的最优密码子数量最少,分析它们的最优密码子数据可知,山楂属 11 个种植物的最优密码子都大多以 A 或 U 作为第三位碱基,说明其最优密码子偏向于使用 A 和 U 作为结尾。对其共有最优密码子进行分析,发现其共有最优密码子有 13 个,分别为 GCA、GCU、AGA、CGA、UGU、CAA、UUA、UUU、AGU、UCU、ACA、UAU 和 GUU,其中有 6 个以 A 作为末碱基,7 个以 U 作为末位碱基,共有密码子的第三位碱基均为 A 和 U。差异密码子有 7 个,分别为 GAC、GAA、GGA、AUA、CUU、AAA 和 ACC,存在差异的最优密码子中,有 4 个以 A 作为第三位碱基,2 个以 C 作为末位碱基,1 个以 U 作为末位碱基。分析山楂属 11 个种植物的最优密码子发现,不存在以 G 作为末位碱基的最优密码子。



A:*C. maximowiczii*;B:*C.kansuensis*;C:*C.oresbia*;D:*C.chungtienensis*;E:*C.rhipidophylla*;F:*C.hupehensis*;G:*C.cuneata*;H:*C.marshallii*;I:*C.pinnatifida*;J:*C.scabrifolia*;K:*C.bretschneideri*。GC_3表示密码子第三位碱基的 G+C 含量,G_{12}表示密码子第一、第二位碱基的 G+C 含量的平均值,R^2表示决定系数。

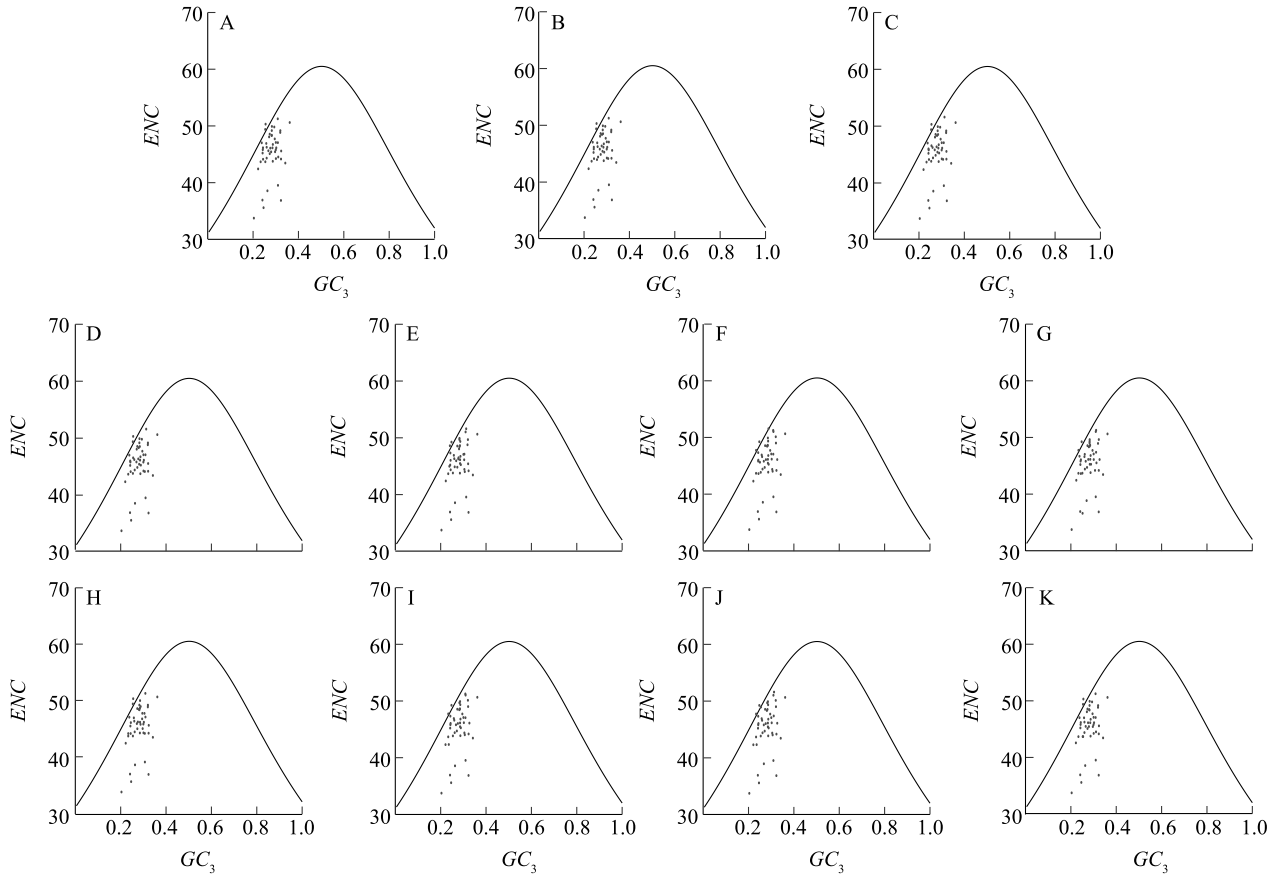
图 5 山楂属植物叶绿体基因组中性绘图

Fig.5 Neutrality plot analysis of chloroplast genomes from *Crataegus* species

表 4 山楂属植物叶绿体基因组最优密码子

Table 4 The optimal codons of chloroplast genomes of *Crataegus*

物种	最优密码子
C. maximowiczii	GCA、GCU、AGA、CGA、GAC、UGU、CAA、GAA、GGA、AUA、CUU、UUA、AAA、UUU、AGU、UCU、ACA、UAU、GUU
C. kansuensis	GCA、GCU、AGA、CGA、GAC、UGU、CAA、GAA、GGA、AUA、CUU、UUA、AAA、UUU、AGU、UCU、ACA、ACC、UAU、GUU
C. oresbia	GCA、GCU、AGA、CGA、GAC、UGU、CAA、GAA、GGA、AUA、CUU、UUA、AAA、UUU、AGU、UCU、ACA、ACC、UAU、GUU
C. chungtienensis	GCA、GCU、AGA、CGA、GAC、UGU、CAA、GAA、GGA、AUA、CUU、UUA、UUU、AGU、UCU、ACA、ACC、UAU、GUU
C. rhipidophylla	GCA、GCU、AGA、CGA、GAC、UGU、CAA、GAA、GGA、AUA、CUU、UUA、UUU、AGU、UCU、ACA、ACC、UAU、GUU
C. hupehensis	GCA、GCU、AGA、CGA、GAC、UGU、CAA、GAA、GGA、CUU、UUA、AAA、UUU、AGU、UCU、ACA、ACC、UAU、GUU
C. cuneata	GCA、GCU、AGA、CGA、GAC、UGU、CAA、GAA、GGA、AUA、UUA、AAA、UUU、AGU、UCU、ACA、ACC、UAU、GUU
C. marshallii	GCA、GCU、AGA、CGA、GAC、UGU、CAA、CUU、UUA、AAA、UUU、AGU、UCU、ACA、ACC、UAU、GUU
C. pinnatifida	GCA、GCU、AGA、CGA、GAC、UGU、CAA、GAA、GGA、AUA、CUU、UUA、AAA、UUU、AGU、UCU、ACA、ACC、UAU、GUU
C. scabrifolia	GCA、GCU、AGA、CGA、UGU、CAA、GAA、AUA、CUU、UUA、AAA、UUU、AGU、UCU、ACA、ACC、UAU、GUU
C. bretschneideri	GCA、GCU、AGA、CGA、UGU、CAA、GAA、GGA、AUA、CUU、UUA、AAA、UUU、AGU、UCU、ACA、ACC、UAU、GUU



A~K 见图 5 注。GC_3 表示密码子第三位碱基 G+C 含量,*ENC* 表示有效密码子数。

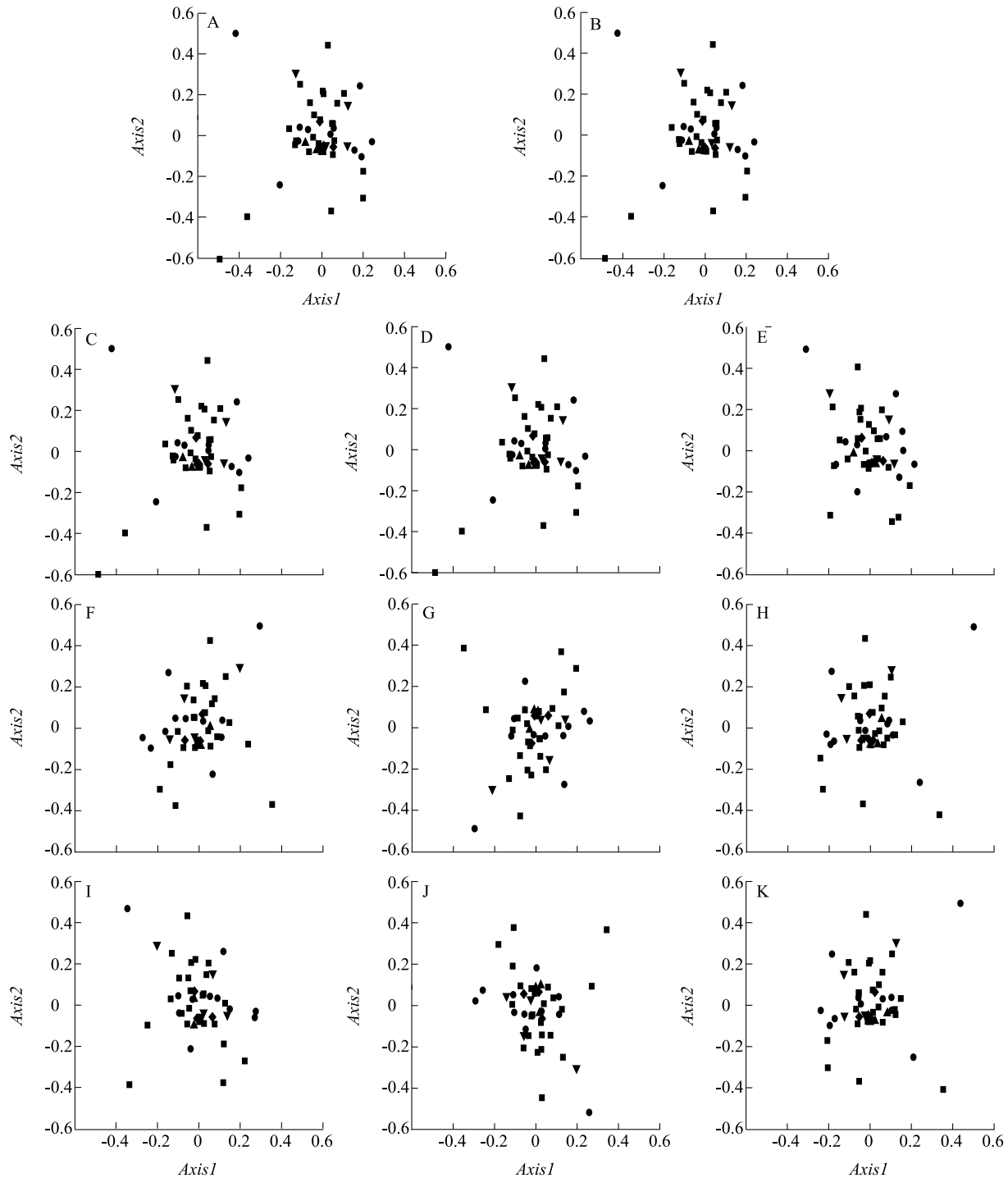
图 6 *ENC*-plot 分析

Fig.6 Analysis of *ENC*-plot

2.2.6 对应性分析 基于 *RSCU* 对山楂属植物叶绿体基因组 48 个共有 CDS 进行对应性分析,结果显示,其第一轴贡献率为11.69%~12.02%,第二轴贡献率为8.78%~8.94%,第三轴贡献率为8.22%~8.37%,第四轴贡献率为7.74%~8.02%,前四轴累计贡献率为36.71%~37.23%,第一轴对变异的贡献率与其他 3 个轴相差较大,为影响其变异的主要因素。为了深入分析其密码子偏好性特征,使用 48 个 CDS 的第一轴和第二轴建立平面坐标系,结果(图 7)显示,山楂属 11 个种植物的 CDS 序列在平面中的分布相似性很高,均显示遗传系统相关基因与保守性开放阅读框的分布相对更加集中,说明这 2 类功能的基因内部存在相似的密码子使用偏好性。而其余 3 种功能的基因分布相对更加分散,说明这 3 种基因的密码子偏好性差异较大。

2.3 系统发育分析

对基于叶绿体 CDS 构建的系统进化树(图 8A)与基于叶绿体全基因组构建的系统发育树(图 8B)进行分析,结果显示,2 种系统发育树具有很高的相似性,*C. kansuensis*、*C. oresbia*、*C. chungtienensis*、*C. bretschneideri*、*C. maximowiczii*、*C. rhipidophylla* 和 *C. marshallii* 在 2 种系统进化树中具有相同的系统发育位置。但 2 种系统发育树也显现出了一定的差异,基于叶绿体 CDS 构建的系统发育树显示 *C. scabrifolia* 被单独归为一个远缘分支,显示其与另外 10 个种的亲缘关系较远;基于叶绿体全基因组序列构建的系统发育树(图 8B)则将 *C. cuneata* 单独归为一个远缘分支。除此之外,基于叶绿体 CDS 构建的系统发育关系显示,*C. hupehensis* 与 *C. pinnatifida* 亲缘关系密切,聚为一类,而基于叶绿体全基因组构建的系统发育树则为 *C. hupehensis*、*C. pinnatifida* 和 *C. scabrifolia* 聚为一支。总的来说,叶绿体基因组的 2 种系统发育树展现出来的系统发育关系既存在着部分差异,也存在着一定的相似性。



▲保守开放性阅读框；■光合作用系统相关基因；●核糖体蛋白相关基因；▼其他蛋白质的编码基因；◆遗传系统相关基因

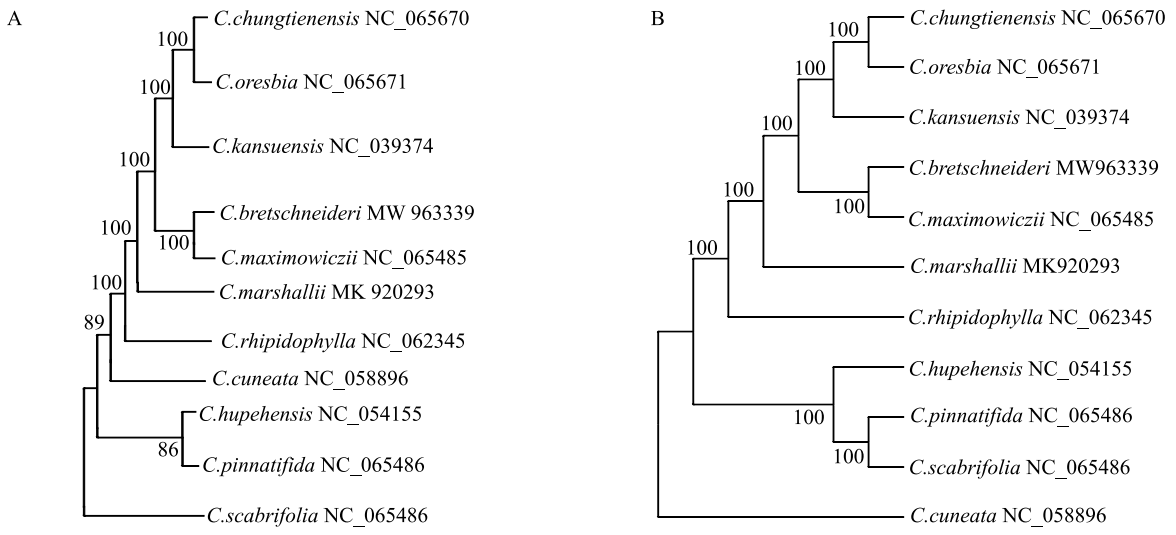
A～K 见图 5 注。*Axis1* 表示第一向量轴，*Axis2* 表示第二向量轴。

图 7 基于 *RSCU* 的基因对应性分析

Fig.7 Gene correspondence analysis based on *RSCU*

3 讨论与结论

植物叶绿体全基因组长度大多为120～200 kb，包含植物体丰富的遗传学信息[35]。随着第二代高通量测序技术的发展和测序成本的降低，已有大量的叶绿体基因组数据被上传至 GenBank 公共数据



A:基于叶绿体基因组 CDS 序列的系统进化树;B:基于叶绿体全基因组序列的系统进化树。

图 8 基于 CDS 序列和叶绿体全基因组构建的山楂属物种系统发育树

Fig.8 Phylogenetic tree of *Crataegus* constructed based on CDS and complete chloroplast genome

库,为植物的系统发育和分子标记研究提供了重要的参考。本研究对山楂属 11 个种的植物叶绿体基因组进行了系统发育与密码子偏好性分析,对于深入研究山楂属植物的进化关系具有一定的意义。本研究选取了山楂属 11 个种的植物叶绿体基因组进行分析,结果显示,山楂属植物的叶绿体基因组结构保守,叶绿体基因组长度变异较小,未发现任何基因组倒置和重排现象,这与悬钩子属植物叶绿体基因组的情况相似[36],但在樟科植物的研究中发现,IR 区存在着部分基因重排现象[37],这与本研究的结果存在一定的差异。重复序列包含植物体的重要进化信息,是控制植物体生长发育的重要部分,重复序列的差异会对植物的遗传发育产生重要影响[38],对所选取的山楂属植物的离散重复序列进行分析,发现正向重复序列、回文重复序列、反向重复序列 3 种离散重复序列在山楂属 11 个种植物中均有分布,而互补重复序列在 *C. maximowiczii* 与 *C. bretschneideri* 中并未检测出,推断 *C. maximowiczii* 与 *C. bretschneideri* 在系统发育关系上可能存在着一定的相似性,这种推断与本研究中 2 种系统进化树展现的系统发育关系也相吻合。

分子进化中性理论认为,基因的碱基突变对密码子的影响是中性的或近似中性的[39]。但如果基因组的密码子受到外界环境选择的影响,则会导致密码子的使用和碱基组成出现偏向性[40]。本研究中选取的山楂属植物叶绿体基因组密码子的 GC_{12} 与 GC_3 的相关系数为0.324~0.525,相关性均未达到显著水平($P>0.05$),GC_{12} 与 GC_3 之间相关性较弱,山楂属植物叶绿体基因组密码子的第一、第二位碱基与第 3 位碱基差异较大,说明选择压力对其密码子有着非常大的影响,而 ENC-plot 和 PR2-plot 绘图分析结果也表明,山楂属植物叶绿体基因组的密码子受选择压力的影响较大。综合以上分析可以看出,本研究中的山楂属植物密码子使用受自然选择因素的影响远大于碱基突变,而影响密码子使用偏好性的主要因素在不同植物物种中也可能存在差异。对应性分析结果显示,遗传系统相关基因与保守性开放阅读框 2 种功能的基因呈现出相似的密码子使用偏性,而其余 3 种功能基因的密码子偏好性存在较大差异,推测这 3 种功能基因的密码子偏好性可能受到多种因素的共同影响。另外,本研究在山楂属 11 个物种中筛选得到17~20 个最优密码子,在这 11 个物种中,均以 A、U 作为结尾的最优密码子数量最多,这一结果与乌头属植物[41]和睡莲属植物[42]的情况相似。分析其共有密码子发现,其共有最优密码子有 13 个,且均以 A 和 U 作为结尾,所有物种中均未发现以 G 作为末位碱基的最优密码子。最优密码子的筛选结果可以为后续山楂属植物的遗传育种工作提供重要的参考依据。

基于 CDS 和叶绿体全基因组构建的 2 种系统

发育树展现出来的系统发育关系具有相似性,这也说明了编码蛋白质氨基酸序列的碱基突变与生物的进化历程存在一定联系,基于叶绿体基因组 CDS 的系统发育关系能在一定程度上对物种的系统发育关系和生物进化历程进行补充。但 *C. cuneata*、*C. hupehensis*、*C. pinnatifida* 和 *C. scabrifolia* 在 2 种系统发育树中的位置存在一定的差别,推测可能是其存在较为特殊的生物进化历程或非编码区碱基序列存在较大差异所导致的。此外,本研究也对山楂属植物的简单重复序列进行了鉴定和分析,可以为后续山楂属植物的分子标记研究提供一定的参考。总之,本研究对山楂属 11 个种植物的叶绿体基因组特征、密码子偏好性及系统发育关系进行了分析,对后续山楂属植物密码子优化、基因组改造以及探索其系统进化关系和增加外源基因表达量具有重要的参考价值。

本研究使用生物信息学手段,对山楂属植物叶绿体基因组进行分析,发现山楂属植物叶绿体基因组结构保守,未发现基因倒置和重排现象,边界扩张收缩幅度小,长度变异保守。对其简单重复序列与离散重复序列进行鉴定,重复序列的种类和数量存在一定的差异。对其密码子偏好性进行分析,结果显示,选择压力均对其密码子偏好性产生深刻的影响,筛选得到的最优密码子数量为17~20 个,使用山楂属 11 个种植物的叶绿体全基因组和 CDS 分别构建系统发育树,发现这 2 种山楂属系统发育树展现出的系统发育关系存在一定相似性。

参考文献:

[1] 费开伟.读山楂种质资源专著——《中国果树志·山楂卷》[J].园艺学报,1998 (1):103.

[2] DEKIC V, RISTIC N, DEKIC B, et al. Phenolic and flavonoid content and antioxidant evaluation of hawthorn (*Crataegus monogyna* Jacq.) fruits and leaves extracts[J]. Bulletin of Natural Sciences Research, 2020, 10(1): 20-25.

[3] LISTON A, WEITEMIER K A, LETELIER L, et al. Phylogeny of *Crataegus* (Rosaceae) based on 257 nuclear loci and chloroplast genomes: evaluating the impact of hybridization[J]. PeerJ, 2021, 9: e12418.

[4] CHEN X L, ZHOU J G, CUI Y X, et al. Identification of *Ligularia* herbs using the complete chloroplast genome as a super-barcode [J]. Frontiers in Pharmacology, 2021, 9: 695.

[5] LI J, LI H Y, ZHI J K, et al. Codon usage of expansin genes in *Populus trichocarpa*[J]. Current Bioinformatics, 2017, 12(5): 452-461.

[6] MORALES-BRIONES D F, KADEREIT G, TEFARIKIS D T, et al. Disentangling sources of gene tree discordance in phylogenomic data sets: testing ancient hybridizations in Amaranthaceae sl[J]. Systematic Biology, 2021, 70(2): 219-235.

[7] CHAKRABORTY S, YENGKHOM S, UDDIN A. Analysis of codon usage bias of chloroplast genes in *Oryza* species[J]. Planta, 2020, 252(4): 1-20.

[8] TANG D F, WEI F, CAI Z Q, et al. Analysis of codon usage bias and evolution in the chloroplast genome of *Mesona chinensis* Benth [J]. Development Genes and Evolution, 2021, 231(1): 1-9.

[9] 王存堂,李子钰,张福娟,等.山楂属果实不同组织乙醇提取物的抗氧化成分及性能研究[J].食品与发酵工业,2021,47(16):117-122.

[10] AIERKEN A, BUCHHOLZ T, CHEN C, et al. Hypoglycemic effect of hawthorn in type II diabetes mellitus rat model[J]. Journal of the Science of Food and Agriculture, 2017, 97(13), 4557-4561.

[11] MIN Q, BAI Y T, ZHANG Y C, et al. Hawthorn leaf flavonoids protect against diabetes-induced cardiomyopathy in rats via PKC-α signaling pathway[J]. Evidence-Based Complementary and Alternative Medicine, 2017. https://doi.org/10.1155/2017/2071952.

[12] 张浣悠,邓秩童,黄嘉泳,等.山楂黄酮的保健功效及提取工艺研究进展[J].食品研究与开发,2021,42(12):212-217.

[13] 张 枭,杜 潇,孙馨宇,等.利用 SSR 标记构建部分山楂资源的基因身份证[J].沈阳农业大学学报,2021,52(2):153-159.

[14] WU X E, LUO D L, ZHANG Y M, et al. Comparative genomic and phylogenetic analysis of chloroplast genomes of hawthorn (*Crataegus* spp.) in southwest China[J]. Frontiers in Genetics, 2022, 13. https://doi.org/10.3389%2Ffgene.2022.900357.

[15] WU L W, CUI Y X, WANG Q, et al. Identification and phylogenetic analysis of five *Crataegus* species (Rosaceae) based on complete chloroplast genomes[J]. Planta, 2021, 254(1): 1-12.

[16] TAI T H, TANKSLEY S D. A rapid and inexpensive method for isolation of total DNA from dehydrated plant tissue[J]. Plant Molecular Biology Reporter, 1990, 8(4): 297-303.

[17] JIN J J, YU W B, YANG J B, et al. GetOrganelle: a fast and versatile toolkit for accurate *de novo* assembly of organelle genomes [J]. Genome Biology, 2020, 21(1): 1-31.

[18] WICK R R, SCHULTZ M B, ZOBEL J, et al. Bandage: interactive visualization of *de novo* genome assemblies[J]. Bioinformatics, 2015, 31(20): 3350-3352.

[19] SHI L C, CHEN H M, JIANG M, et al. CPGAVAS2, an integrated plastome sequence annotator and analyzer[J]. Nucleic acids research, 2019, 47(W1): 65-73.

[20] KEARSE M, MOIR R, WILSON A, et al. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data[J]. Bioinformatics, 2012, 28 (12): 1647-1649.

[21] 王 瑞,欧金梅,李 昕,等.基于简单重复序列标记的药用梅

品种的身份证构建[J].安徽中医药大学学报,2020,39(6):62-67.

[22] BEIER S, THIEL T, MüNCH T, et al. MISA-web: a web server for microsatellite prediction[J]. Bioinformatics, 2017, 33(16): 2583-2585.

[23] KURTZ S, SCHLEIERMACHER C. REPuter: fast computation of maximal repeats in complete genomes[J]. Bioinformatics (Oxford, England), 1999, 15(5): 426-427.

[24] XIA E H, TONG W, WU Q, et al. Tea plant genomics: achievements, challenges and perspectives[J]. Horticulture research, 2020, 7. https://doi.org/10.1038/s41438-019-0225-4.

[25] 梁凤萍,文祥宁,高赫一,等.菊科植物叶绿体基因组特征分析[J].基因组学与应用生物学,2018,37(12):5437-5447.

[26] DARLING A C, MAU B, BLATTNER F R, et al. Mauve: multiple alignment of conserved genomic sequence with rearrangements[J]. Genome Research, 2004, 14(7): 1394-1403.

[27] WALDVOGEL A M, PFENNINGER M. Temperature dependence of spontaneous mutation rates[J]. Genome Research, 2021, 31(9): 1582-1589.

[28] DE OLIVEIRA J L, MORALES A C, Hurst L D, et al. Inferring adaptive codon preference to understand sources of selection shaping codon usage bias[J]. Molecular Biology and Evolution, 2021,38(8):3247-3266.

[29] XING Y P, XU L, CHEN S Y, et al. Comparative analysis of complete chloroplast genomes sequences of *Arctium lappa* and *A.tomentosum*[J]. Biologia Plantarum, 2019,63(1): 565-574.

[30] DUAN H R, ZHANG Q, WANG C M, et al. Analysis of codon usage patterns of the chloroplast genome in *Delphinium grandiflorum* L. reveals a preference for AT-ending codons as a result of major selection constraints[J]. PeerJ, 2021, 9:e10787.

[31] KATOH K, STANDLEY D M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability[J]. Molecular Biology and Evolution, 2013, 30(4): 772-780.

[32] CAPELLA-GUTIÉRREZ S, SILLA-MARTíNEZ J M, GABALDóN T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses[J]. Bioinformatics, 2009,25(15):1972-1973.

[33] LI W, ZHANG C P, GUO X, et al. Complete chloroplast genome of *Camellia japonica* genome structures, comparative and phylogenetic analysis[J]. PLoS One, 2019, 14(5): e0216645.

[34] MENSAH R A, SUN X, CHENG C, et al. Analysis of codon usage pattern of banana basic secretory protease gene[J]. Plant Diseases and Pests, 2019, 10(1): 1-9.

[35] 童一涵,郑 倩,杜新明,等. 多齿红山茶叶绿体基因组序列特征分析[J]. 植物资源与环境学报,2022,31(5):27-36.

[36] 于丽平,孙孟涛,贺志敏,等. 川莓和峨眉悬钩子叶绿体比较基因组学及其系统发育关系分析[J].分子植物育种, 2022. http://kns.cnki.net/kcms/detail/46.1068.S.20220729.1007.004.html.

[37] 田永靖. 樟科植物比较叶绿体基因组与系统发育研究[D]. 南京: 南京大学, 2021.

[38] KELLER J, ROUSSEAU-GUEUTIN M, MARTIN G E, et al. The evolutionary fate of the chloroplast and nuclear *rps16* genes as revealed through the sequencing and comparative analyses of four novel legume chloroplast genomes from Lupinus[J]. DNA Research, 2017, 24(4): 34.

[39] ZHANG R Z, ZHANG L, WANG W, et al. Differences in codon usage bias between photosynthesis-related genes and genetic system-related genes of chloroplast genomes in cultivated and wild *solanum* species[J]. International Journal of Molecular Sciences, 2018, 19(10): 3142.

[40] LIU H B, LU Y Z, LAN B L, et al. Codon usage by chloroplast gene is bias in *Hemiptelea davidii*[J]. Journal of Genetics, 2020, 99(1): 1-11.

[41] 樊东昌,穆赢通,贾俊英,等. 乌头属药用植物叶绿体基因组密码子特征和系统发育分析[J].分子植物育种, 2022. http://kns.cnki.net/kcms/detail/46.1068.S.20220711.1339.002.html.

[42] 毛立彦,黄秋伟,龙凌云,等. 7 种睡莲属植物叶绿体基因组密码子偏好性分析[J].西北林学院学报,2022,37(2):98-107.

(责任编辑:陈海霞)