红花龙胆叶绿体基因组特征及适应性进化分析

邓 港[1,2], 吴田泽[1,2], 高冉冉[1], 王梦月[1], 刘 霞[2*], 向 丽[1*]

(1. 中国中医科学院中药研究所, 中药鉴定与安全性评估北京市重点实验室, 北京 100700; 2. 武汉理工大学化学化工与生命科学学院, 湖北 武汉 430070)

摘要: 红花龙胆 (*Gentiana rhodantha*) 为特色苗族药材, 在治疗湿热黄疸、小便不利、肺热咳嗽等方面具有显著疗效, 但在物种进化关系及分类鉴定方面存在争议。本研究对红花龙胆叶绿体基因组进行了二代、三代测序, 并分析其结构特点及适应性进化分析。结果表明, 红花龙胆叶绿体基因组全长148 844 bp, 大单拷贝区 (LSC) 80 076 bp、小单拷贝区 (SSC) 17 596 bp和反向重复区 (IR) 25 586 bp, GC含量37.75%。共注释到124个基因, 包括80个蛋白编码基因 (CDS)、36个tRNA基因和8个rRNA基因; 红花龙胆叶绿体基因组密码子偏好性较弱, 影响因素主要是自然选择, 最优密码子为CUU、UCU、UCA、CCA、ACU; MISA共发现169个SSR, 其中单核苷酸重复最多 (114个, 67.50%), 其次是二核苷酸重复 (43个, 25.44%); 与同组及同属其他物种相比, 红花龙胆叶绿体基因的Ka/Ks值基本小于1, 表明在长期的进化过程中受到了较强的纯化选择, 存在进化事件的光合作用基因*psaI*和表达相关基因*rpl22*、*rps11*出现组间差异, 且系统发育分析结果支持红花龙胆与龙胆属狭蕊组聚为一支, 并与其他组能够明显区别开, 支持了狭蕊组独立成属的观点。本研究将为今后开展红花龙胆叶绿体基因工程、分子育种等研究提供参考依据。

关键词: 红花龙胆; 叶绿体基因组; 系统发育; 适应性进化; 密码子; 简单重复序列

中图分类号: R931 **文献标识码**: A **文章编号**: 0513-4870(2022)10-3240-14

Characteristics and adaptive evolution analysis of the chloroplast genome of *Gentiana rhodantha*

DENG Gang[1,2], WU Tian-ze[1,2], GAO Ran-ran[1], WANG Meng-yue[1], LIU Xia[2*], XIANG Li[1*]

(*1. Key Laboratory of Beijing for Identification and Safety Evaluation of Chinese Medicine, Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing 100700, China; 2. School of Chemistry, Chemical Engineering and Life Sciences, Wuhan University of Technology, Wuhan 430070, China*)

Abstract: *Gentiana rhodantha* is a characteristic medicinal material of Miao Ethnomedicine. It has significant curative effect in the treatment of acute jaundice hepatitis, dysentery, pediatric pneumonia and bronchitis, etc. However, the evolutionary relationship and taxonomic identification of *G. rhodantha* are controversial. In this study, we sequenced the chloroplast genome of *G. rhodantha* using the second and third generation sequencing technology. Then, the structural characteristics and suitability evolution characteristics were analyzed. The results showed that the *G. rhodantha* chloroplast genome was 148 844 bp in length with 37.75% GC content, consisting of a large single copy region (LSC) of 80 076 bp, a small single copy region (SSC) of 17 596 bp and an inverted repeat region (IR) of 25 586 bp. A total of 124 genes were annotated, including 80 protein-coding genes, 36 tRNA genes, and 8 rRNA genes; the chloroplast genome of *G. rhodantha* has a weak codon preference, and the influencing factors are mainly natural selection. The optimal codons are CUU, UCU, UCA, CCA, and ACU. A total of 169

收稿日期: 2022-05-05; 修回日期: 2022-05-30.

基金项目: 国家自然科学基金委−贵州喀斯特中心项目 (U1812403-1); 国家重点研发计划资助项目 (2019YFC1711100).

*通讯作者 E-mail: lxiang@icmm.ac.cn; lrx1125@126.com

DOI: 10.16438/j.0513-4870.2022-0537

SSRs were found in MISA, of which the single nucleotide repeats were the most (114, 67.50%), followed by dinucleotide repeats (43, 25.44%). The phylogenetic analysis support that *G. rhodantha* belong to *Sect. Stenogyne* which can be clearly distinguished from other groups. Compared with other species, the Ka/Ks value of chloroplast genes of *G. rhodantha* is basically less than 1 except for *psaI*, *rpl22* and *rps11*, indicating that they have been subjected to strong purification selection in the long-term evolutionary process. The photosynthesis gene *psaI* and the expression-related genes *rpl22* and *rps11* showed differences between groups, which supported the view that *Sect. Stenogyne* was an independent genus. This study will provide a reference for future researches on chloroplast genetic engineering and molecular breeding of *G. rhodantha*.

Key words: *Gentiana rhodantha*; chloroplast genome; phylogeny; adaptive evolution; codon; simple sequence repeat

红花龙胆 (*Gentiana rhodantha*) 为龙胆科龙胆属药用植物, 别名红龙胆、雪理梅、小青鱼胆、星秀花、龙胆草[1]等, 为2015版《中国药典》[2]中新增中药材品种, 2020年版中也有收载, 红花龙胆具有清热除湿、解毒泻火、止咳的功效, 主治湿热黄疸、小便不利、肺热咳嗽等症。与龙胆科其他药用植物以环烯醚萜类成分为主不同, 红花龙胆中主要成分为黄酮类物质芒果苷(mangiferin), 而环烯醚萜苷类及三萜类化合物等的含量甚微[3]。Xu等[4]对从红花龙胆中分离得到的多种化合物进行检测后发现, 芒果苷具有乙酰胆碱酯酶抑制活性。Yao[5]通过对不同产地红花龙胆药材进行DPPH抗氧化研究, 发现叶、花和地上部分具有较好的抗氧化活性。

红花龙胆分布于黄河以南多个省区[6], 云南、贵州等省份的蕴藏量相对较多, 且在贵州喀斯特地区分布范围较广, 九个地州市均有分布, 但野生资源贮量较小[7]。随着贵州苗药制药厂家的大量收购, 加上近年来旅游业的发展, 人为破坏的加剧, 使得野生红花龙胆储量正在急剧减少, 资源日趋匮乏, 已远不能满足苗药市场需求, 野生资源保护已迫在眉睫[8]。目前已经开发了多种与红花龙胆相关的复方配伍剂型, 如肺力咳合剂[9]、康复灵片和莲龙胶囊, 市场对红花龙胆的需求呈逐年上升趋势。龙胆属约400种, 属下有11个组, 我国分布247种, 大多数种类集中在西南山岳地区, 主要生长在高山流石滩、高山草甸和灌丛中[10]。该属植物生境多为高山和高海拔地区, 生态系统脆弱, 对气候变化敏感。已有研究显示, 不同龙胆属植物所能耐受的气候变化幅度并不一致, 气候持续变暖可能增加部分种类的灭绝风险[11]。红花龙胆作为龙胆属狭蕊组的代表物种, 生境立体气候特征显著, 地形地貌复杂, 是研究西南龙胆属植物环境与进化关系的理想物种之一。

不同于核基因组及线粒体基因组, 叶绿体基因组多为母系遗传。由于叶绿体基因组结构、长度和基因种类的保守性[12], 其编码区和非编码区进化速率与模式具有物种鉴别意义[13], 可用于解决复杂类群的系统发育、DNA分子标记筛选等问题, 如今已广泛用于叶绿体基因工程[14]。近年来叶绿体基因组的数据信息被大量利用, 如适应性进化分析等。基因的结构与功能息息相关, 分析叶绿体基因组的适应性进化对于研究基因的结构变化和功能变异有着深远影响。不同植物叶绿体基因组中存在着不同类型的变异, 且各个编码基因的进化速率并不相同, 这可用于探究复杂类群的系统发育问题, 为植物种属进化关系提供依据。

本研究以红花龙胆为材料, 通过高通量测序、组装和基因注释, 得到红花龙胆叶绿体全基因组, 并完成叶绿体基因组系统发育分析、密码子偏好性分析、重复序列分析、进化压力分析, 本研究为今后开展红花龙胆叶绿体基因工程、遗传多样性分析、分子育种等研究提供参考依据。

材料与方法

植物材料和DNA提取　红花龙胆样品于2021年7月27日采集于贵州省黔东南州东南部雷公山, 经纬度为105°36′15″N, 24°59′3″E; 将植株健康的、新鲜的叶部位以锡箔纸包裹后置于液氮中, 于−80 ℃保存备用。经贵州中医药大学王波鉴定为红花龙胆, 标本存放于中国中医科学院中药研究所, 标本号为zxht01-龙2-6。总DNA提取采用改良的CTAB方法, 从干燥叶片中提取植物总DNA。经检测合格后, 用于文库构建。

基因组测序、组装和注释　文库检测合格后, 按照有效浓度及目标下机数据量的需求将不同文库混合至Flowcell芯片中, cBOT成簇后使用高通量测序平台Illumina NovaSeq进行测序。建库完成后将一定浓度和体积的DNA文库加入到Flowcell中, 并将Flowcell转移到Oxford Nanopore PromethION测序仪进行三代测序, 测序工作由武汉贝纳科技服务有限公司完成。对测序得到的序列进行低质量数据过滤得到有效数据, 采用合适的数据量进行组装及后续分析。使用

Flye 软件 (version: v.2.8.3; 参数: –meta –plasmids)[15] 进行基因组拼接, 采用有参拼接, 参考基因组为NCBI已有的红花龙胆序列NC050307。将拼接结果与参考基因组进行Blastn比对, 基于序列测序深度、读长比对情况以及与近缘物种的比对情况等, 确定叶绿体的基因组连接关系。三代组装后, 利用二代数据进行纠错与修正, 以获得错误率相对较低的结果。获得完整序列后, 利用针对叶绿体的注释软件CPGAVAS2[16]进行基因注释和圈图绘制, 选择的参照基因组为NCBI已有红花龙胆叶绿体基因组NC050307, 并在Geneious[17]软件中手动校正注释结果, 叶绿体基因组功能注释包括编码基因预测和非编码RNA注释 (rRNA和tRNA注释)。将组装并注释完成后的红花龙胆叶绿体全基因组序列提交至GenBank数据库, NCBI检索号ON378800。

系统发育树的构建 得到红花龙胆叶绿体全基因组后, 从NCBI数据库 (https://www.ncbi.nlm.nih.gov/) 检索到已公布的9条龙胆属物种序列, 依据中国植物志的分类系统, 分为以下7组: 狭蕊组 (条纹龙胆 *G. striata*、高贵龙胆 *G. gentilis*)、秦艽组 (秦艽 *G. macrophylla*)、微籽组 (微籽龙胆 *G. delavayi*)、龙胆草组 (龙胆 *G. scabra*、条叶龙胆 *G. manshurica*)、高山龙胆组 (太白龙胆 *G. apiata*)、匐茎组 (矮龙胆 *G. wardii*)、耳褶龙胆组 (耳褶龙胆 (*G. otophora/Kuepferia otophora*)。选择夹竹桃科 (长春花 *Catharanthus roseus*) 和龙胆科 (扁蕾 *Gentianopsis barbata*) 作为亲缘关系较近的外类群, 利用MAFFT[18]软件进行12种植物的序列多重对比, 序列检查发现微籽龙胆 *G. delavayi* 有Y存在, 导致建树时运行错误, 根据Y=C/T的规则, 手动将Y替换。使用MEGA7[19]软件利用邻接法 (NJ) 建立系统发育树 (模型: p-distance; bootstrap method; 1 000 replications), 并使用IQtree[20]软件利用最大似然法 (ML) 构建系统发育树 (核苷酸替换模型选择GTR+G; 系统发育树各分支的bootstrap values通过进行1 000次自展重复分析获得), 将建树结果合并, 最后使用FigTree软件对系统发育树进行可视化展示。

密码子偏好性分析 通过CPGAVAS2注释红花龙胆叶绿体基因组得到蛋白编码序列 (CDS), 为了避免样本误差和数据冗杂, 手动筛选, 并移除其中的重复基因序列以及长度小于300 bp的编码序列, 将符合分析条件的CDS用于后续分析。使用CUSP在线工具对各CDS的密码子第1、2、3位核苷酸上的GC含量进行在线分析, 利用CodonW对各个基因的密码子在第3核苷酸上的A、G、C、T含量进行计算; 同时利用该软件对各基因的氨基酸长度 (Laa)、有效密码子数 (ENC)、同义密码子相对使用度 (RSCU) 及最优密码子使用频率 (FOP) 进行计算; 使用R语言中的ggplot2和aplot安装包进行绘图, 对不同氨基酸的密码子RSCU值进行可视化, 并将所得到的数据使用Microsoft Excel软件进行统计及变量间的相关性分析。中性绘图分析: 取各基因GC1及GC2的平均值, 记为GC12, 以各基因的GC12为纵坐标、GC3为横坐标绘制散点图, 并对二者的相关性进行分析。ENC-plot分析: 取各基因的ENC为纵坐标、GC3为横坐标绘制散点图, 根据有效密码子ENC=2+GC3+29/[GC32+(1-GC3)2] 的公式计算各基因的理论ENC值, 并以GC3为横坐标、理论ENC值为纵坐标绘制标准曲线。PR2-plot分析: 以G3/(G3+C3)和A3/(A3+T3)分别为横纵坐标绘制散点图, 对密码子第3位核苷酸上的碱基组成情况进行分析, 从而探讨突变和自然选择对密码子使用偏好性的影响。最优密码子分析: 将CDS按ENC值由高至低排序, 从两端各选出10%的基因数作为高、低表达库, 利用codonW软件运行得到两个库中编码氨基酸密码子的RSCU。

重复序列分析 采用MISA[21]软件对叶绿体基因组进行SSR检测, 其参数设置如下: 对应的各个重复单元 (unit size) 的最少重复次数分别为: 1～8、2～4、3～4、4～3、5～3、6～3。使用perl脚本处理MISA输出的结果, 确定SSR在红花龙胆叶绿体基因组中的具体位置。

选择压力分析 通过非同义替换位点替换次数 (Ka) 与同义替换位点替换次数 (Ks) 的比值 (Ka/Ks) 判断红花龙胆与同属同组的条纹龙胆、同属异组的粗茎秦艽和条叶龙胆、同科混伪品扁蕾、外类群长春花之间叶绿体蛋白编码基因是否存在选择压力。首先利用CPGAVAS2提取以上物种的CDS基因, 筛选出54条共有CDS基因, 提取的基因序列通过MAFFT进行比对, 然后用DnaSPv5[22]软件计算Ka和Ks值, 统计各基因的Ka/Ks值, 绘制不同功能基因的Ka/Ks图。得到Ka/Ks大于1的CDS基因后, 利用MAFFT软件进行所有12种植物中该基因序列的多重对比。使用MEGA7软件利用邻接法 (NJ) 建立系统发育树 (模型: p-distance; bootstrap method; 1 000 replications), 并通过Geneious可视化碱基组成及组内组间碱基变异情况。

结果与分析

1 红花龙胆叶绿体基本结构

新测序获得的红花龙胆叶绿体基因组与绝大多数被子植物叶绿体基因组一样, 为共价闭合的双链环状分子 (图1), 全长148 844 bp, LSC、SSC与IR的长度分别为80 076 bp、17 596 bp、25 586 bp。全基因组的GC含量为37.75%, 其中GC含量最高的是反向重复区

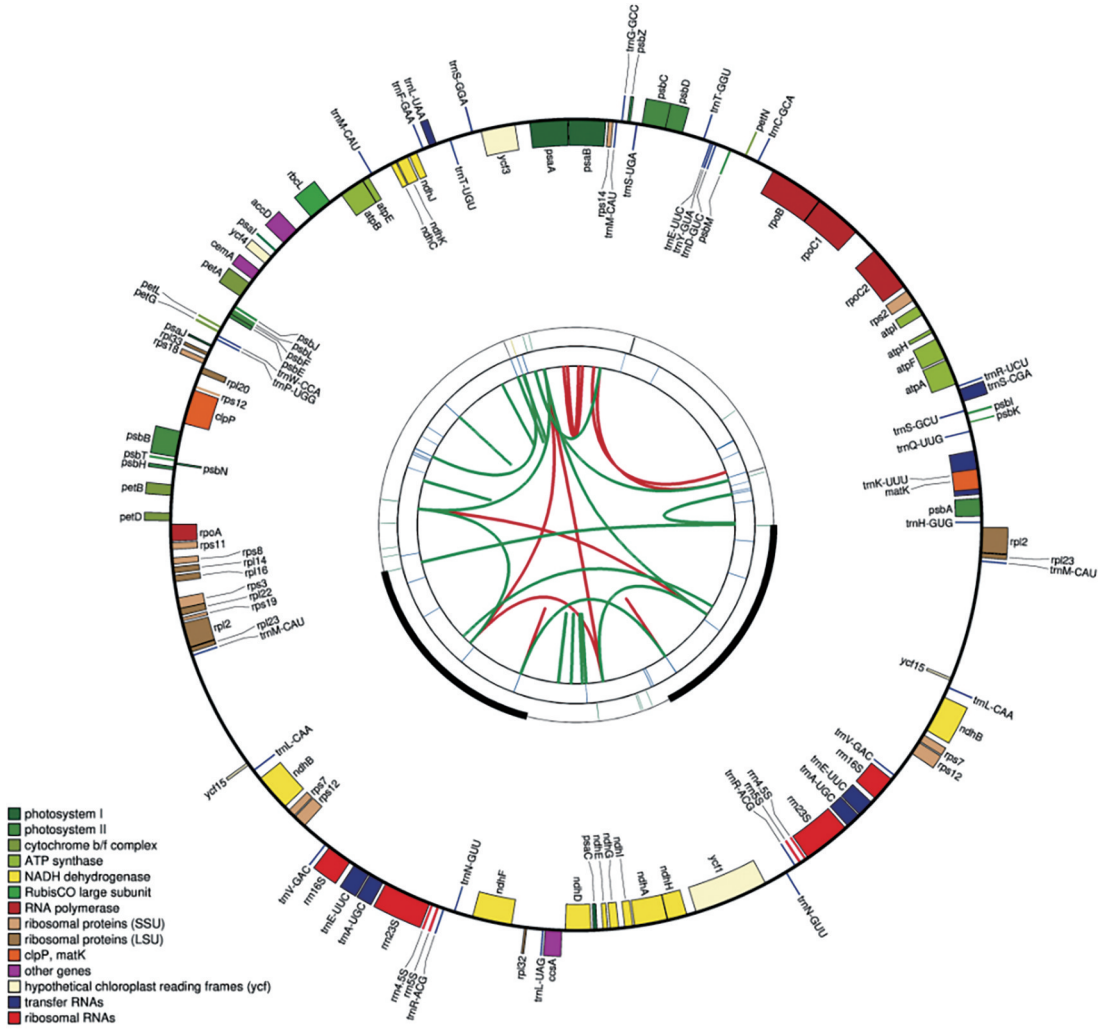

Figure 1　Chloroplast genome map of *G. rhodantha* exported with CPGAVAS2. Genes inside and outside the circle are transcribed in clockwise and counter clockwise direction, respectively. Genes are color-coded based on their functions. The inner circle represents the linear relationship of the genes

(43.47%), 大单拷贝区和小单拷贝区为35.53%和31.23% (表1)。CPGAVAS2结果显示, 红花龙胆叶绿体基因组共注释到124个基因, 包括80个蛋白质编码基因 (CDS)、36个tRNA、8个rRNA, 其中绝大部分为单拷贝基因, 15个基因位于反向重复区 (IR) 为双拷贝基因。57个CDS基因分布于LSC区域、12个分布于SSC区域、11个分布于IR区域。按照功能的差异可以将这些基因分为4个大类 (光合作用相关、自我复制、其他基因、未知功能), 红花龙胆叶绿体基因中与光合作用和自我复制相关的基因占绝大多数 (表2)。

Table 1　Basic features of the chloroplast genome of *G. rhodantha*

Genome feature	Feature value
Genome size/bp	148 844
The length of IR/bp	25 586
The length of LSC/bp	80 076
The length of SSC/bp	17 596
GC content of genome/%	37.75
GC content of IR/%	43.47
GC content of LSC/%	35.53
GC content of SSC/%	31.23
Number of total genes	124
Number of protein-coding genes	80
Number of rRNA genes	8
Number of tRNA genes	36

2　系统发育分析

由于龙胆属物种众多, 分类复杂, 经典分类学观点往往难以准确阐述各物种亲缘关系。基于红花龙胆及其他龙胆属叶绿体全基因组序列, 以长春花、扁蕾为外类群, 使用NJ法和ML法构建系统发育树。结果显示, NJ树与ML树中各类群间的拓扑结构完全一致, 各枝支持率均为100% (图2); 外类群夹竹桃科长春花和龙胆科扁蕾被分出, 而另一大枝为龙胆属各类群。红花龙胆与高贵龙胆亲缘关系最近, 其次为条纹龙胆。

Table 2 Genome classification and quantity of the chloroplast genome of *G. rhodantha*

Gene group	Gene name	Number
ATP synthase	*atpE*, *atpI*, *atpA*, *atpB*, *atpF*, *atpH*	6
Photosystem II	*psbD*, *psbJ*, *psbN*, *psbI*, *psbL*, *psbF*, *psbZ*, *psbT*, *psbE*, *psbM*, *psbK*, *psbB*, *psbC*, *psbH*, *psbA*	15
Rubisco large subunit	*rbcL*	1
RNA polymerase	*rpoB*, *rpoA*, *rpoC2*, *rpoC1*	4
Ribosomal proteins (LSU)	*rpl32*, *rpl23*, *rpl20*, *rpl16*, *rpl22*, *rpl33*, *rpl2*, *rpl14*	8
Other genes	*cemA*, *ccsA*, *accD*	3
Cytochrome b/f complex	*petG*, *petD*, *petB*, *petA*, *petL*, *petN*	6
Transfer RNAs	*trnG-GCC*, *trnR-UCU*, *trnW-CCA*, *trnA-UGC*, *trnS-CGA*, *trnS-GCU*, *trnE-UUC*, *trnC-GCA*, *trnV-GAC*, *trnT-UGU*, *trnP-UGG*, *trnD-GUC*, *trnK-UUU*, *trnM-CAU*, *trnT-GGU*, *trnN-GUU*, *trnY-GUA*, *trnS-GGA*, *trnL-UAG*, *trnL-UAA*, *trnF-GAA*, *trnH-GUG*, *trnR-ACG*, *trnL-CAA*, *trnS-UGA*, *trnQ-UUG*	26
Ribosomal RNAs	*rrn23S*, *rrn5S*, *rrn4.5S*, *rrn16S*	4
Protease	*clpP*	1
Hypothetical chloroplast reading frames	*ycf4*, *ycf1*, *ycf3*, *ycf15*	4
Ribosomal proteins (SSU)	*rps14*, *rps12*, *rps8*, *rps2*, *rps11*, *rps18*, *rps19*, *rps3*, *rps7*	9
NADH dehydrogenase	*ndhG*, *ndhF*, *ndhJ*, *ndhI*, *ndhK*, *ndhH*, *ndhA*, *ndhE*, *ndhD*, *ndhC*, *ndhB*	11
Photosystem I	*psaA*, *psaC*, *psaB*, *psaI*, *psaJ*	5
Maturase	*matK*	1

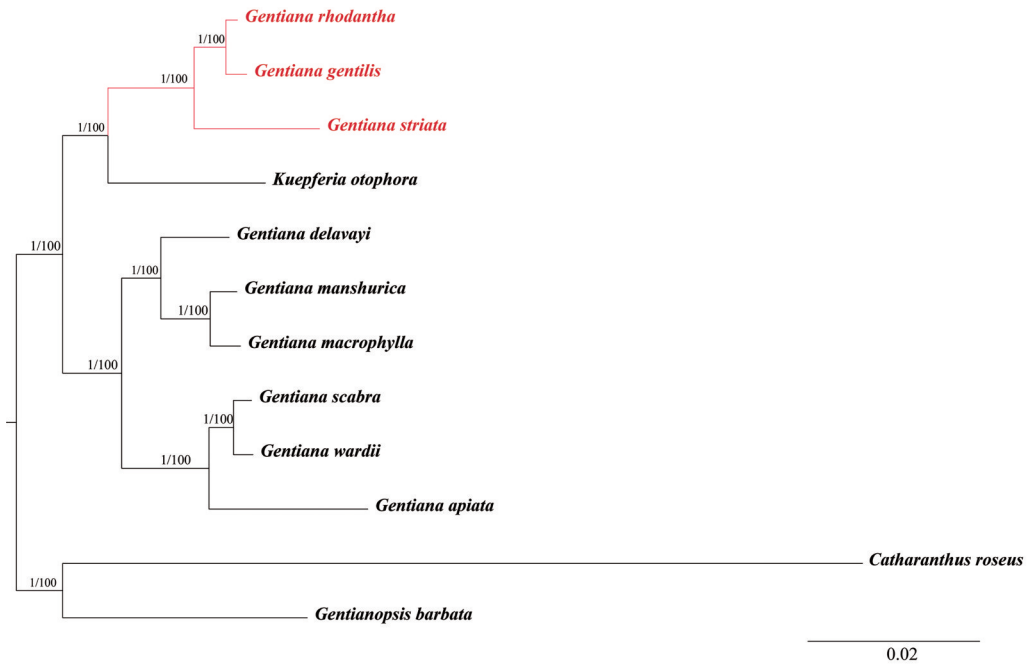

Figure 2 Phylogenetic analysis of *G. rhodantha* and its closely related species. The tree was constructed with 12 chloroplast genomes using the Neighbor-joining method (NJ) by MEGA7 (p-distance; bootstrap method; 1 000 replications), and using Maximum Likelihood method (ML) by IQtree (GTR+G model; 1 000 bootstrap replicates) with *C. roseus* and *G. barbata* serving as the outgroups. Numbers above nodes are support values with NJ on the left and ML bootstrap values on the right. The results of the two methods are consistent

耳褶龙胆与红花龙胆的类群关系相对更近。与红花龙胆亲缘关系相对较远的可分为两大枝, 微籽龙胆、条叶龙胆、秦艽聚为一枝, 而龙胆、矮龙胆、太白龙胆聚为一枝。通过叶绿体全基因组, 红花龙胆可以与其他龙胆属物种及混伪品区别开来。

3 密码子偏好性分析

从红花龙胆叶绿体基因组中得到的80条蛋白质编码序列 (CDS), 经过手动筛选后获得51条符合分析条件的CDS序列, 51条CDS序列的密码子的平均GC含量 (GCall) 为38.19%, 密码子第1位碱基组成的GC含量 (GC1) 为46.05%, 第2位碱基组成的GC含量 (GC2) 为39.91%, 第3位碱基的GC含量 (GC3) 为28.16%, 表明密码子不同位置的GC含量有差异, 其分布频率也有所不同。各位置上的GC含量平均值由高到低依次为第

1 位 > 第 2 位 > 第 3 位, 说明在红花龙胆中, 叶绿体编码基因多为 A/U 碱基, ENC 值为 36.89～56.16, 平均值为46.525, 其中 51 条 CDS 密码子中有 37 条 ENC 值 > 45 (表 3); 可推断出其密码子偏性较弱。通过 CodonW 软件得到 CDS 序列中相对同义密码子使用度 (RSCU), 将每种氨基酸对应的同义密码子 RSCU 值对进行绘图 (图 3); RSCU 分析表明 (表 4), 有 30 个密码子 RSCU > 1, 以 A 和 U 结尾的较多, 其中以 U 结尾的有 16 个, 以 A 结尾的为 13 个, 以 G 结尾的有 1 个, 说明红花龙胆叶绿体基因组倾向于以 A 和 U 结尾, 为偏好密码子; 反之, 以 G、C 结尾的则为非偏好密码子。

3.1　中性绘图分析　突变是密码子使用偏好性的主要影响因素, 若 GC12 和 GC3 的值呈显著相关, 则表明 3 个位点的碱基具有相同的变异模式, 反之则表明变异模式差异较大[23]。中性绘图分析表明 (图 4A), GC12 和 GC3 的取值范围分别为 0.347 8～0.554 0 和 0.207 7～0.392 4, 绝大部分都分布在对角线的上方, GC12 和 GC3 相关系数为 0.085 8, 相关性不显著, 回归系数为 0.107 5, 说明密码子第 1、2 位碱基的突变模式与第 3 位有所不同, 其密码子偏好性受到外界选择下的定向突变压力作用强于自身突变下的非定向突变压力作用, 主要受自然选择影响。

3.2　ENC-plot 绘图分析　红花龙胆叶绿体基因组密码子的 ENC-plot 分析 (图 4B) 显示, 在标准曲线上及周围分布了红花龙胆的少部分基因, 表明了红花龙胆的 ENC 值实际与期望接近, 而在标准曲线下方则有大部分基因, 并离曲线较远, 说明红花龙胆叶绿体基因组密码子偏好性受自然选择影响较大。

3.3　PR2-plot 分析　采用 PR2-plot 绘图分析各编码基因部分氨基酸 A、G 和 T、C 之间的关系, 中心点代表 A=T, C=G, 从中心点向其他位点分布的矢量则显示该基因的偏倚程度及方向[24]。结果显示 (图 4C), 红花龙胆叶绿体基因组中各基因在 4 个象限中分布并不均匀, 大部分基因分布于平面图的下方和右侧, 表明在碱

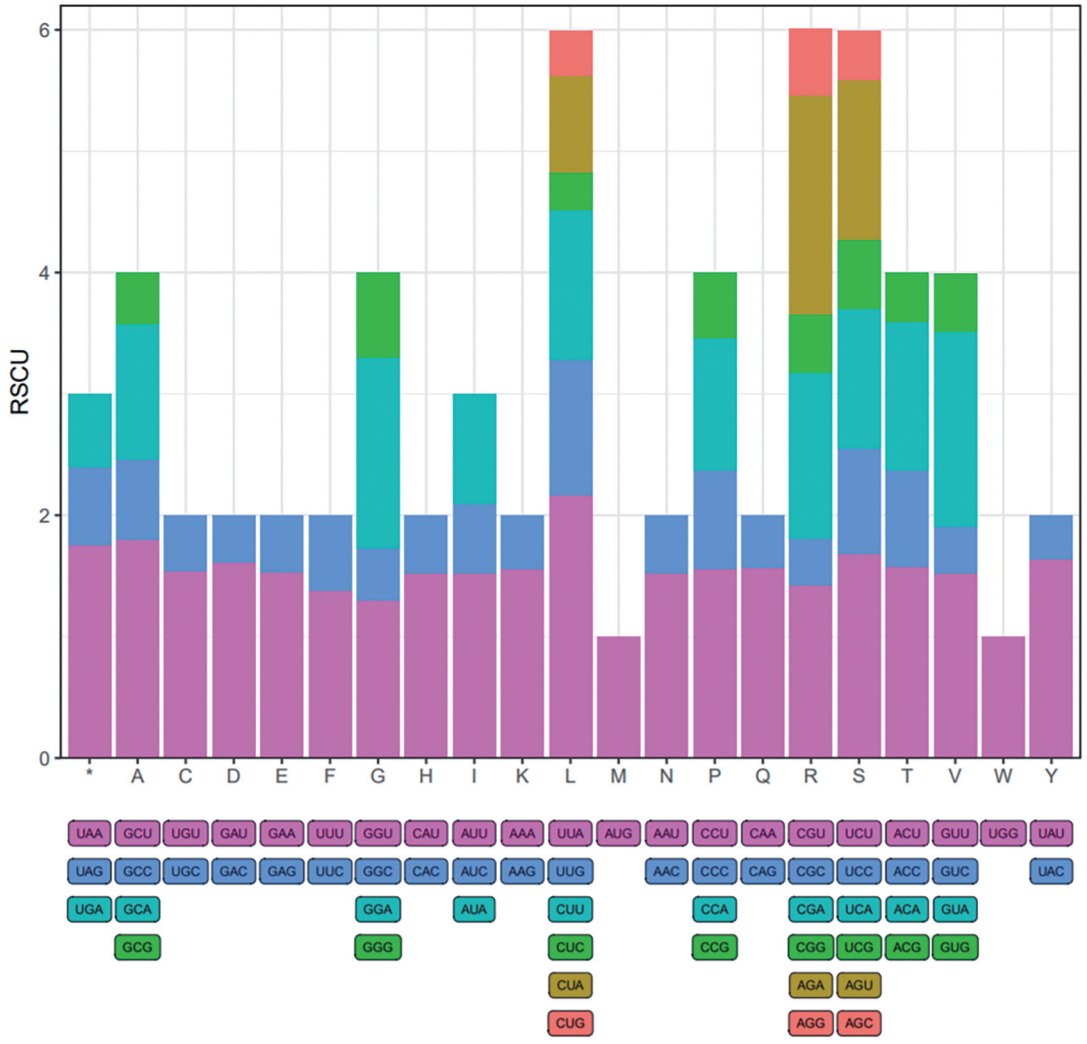

Figure 3　CDS gene relative synonymous codon usage (RSCU). Different codons for the same amino acid are represented by different colors

Table 3 Main parameters in chloroplast genomics of *G. rhodantha*

Gene	GC1	GC2	GC3	GC	ENC	Laa	CAI	CBI	Fop	Gene	GC1	GC2	GC3	GC	ENC	Laa	CAI	CBI	Fop
accD	0.430 2	0.371 5	0.310 1	0.372	46.2	357	0.179	−0.18	0.327	*psbA*	0.491 5	0.432 2	0.299 4	0.409	42	353	0.283	0.14	0.498
atpA	0.555 1	0.387 6	0.279 5	0.412	48.11	507	0.193	−0.053	0.383	*psbB*	0.540 3	0.461 7	0.294 7	0.432	46.47	508	0.192	−0.065	0.377
atpB	0.568 9	0.417 2	0.287 4	0.425	48.24	500	0.191	−0.024	0.395	*psbC*	0.540 1	0.459 9	0.331 2	0.444	46.85	473	0.182	−0.028	0.393
atpE	0.541 4	0.398 5	0.263 2	0.404	49.82	132	0.151	−0.118	0.336	*psbD*	0.519 8	0.437 9	0.316 4	0.426	44.11	353	0.254	0.056	0.45
atpF	0.459	0.316 9	0.311 5	0.364	49.03	182	0.151	−0.125	0.337	*rbcL*	0.578 3	0.432 2	0.302 7	0.439	47.61	478	0.259	0.047	0.448
atpI	0.490 4	0.384 6	0.245 2	0.374	41.14	207	0.181	−0.033	0.377	*rpl14*	0.552 8	0.374	0.252	0.396	45.91	122	0.186	−0.004	0.398
ccsA	0.352 8	0.349 7	0.263 8	0.322	48.58	325	0.138	−0.216	0.282	*rpl16*	0.523 1	0.523 1	0.207 7	0.418	36.89	130	0.116	−0.143	0.333
cemA	0.407 9	0.285 1	0.302 6	0.332	52.85	227	0.203	−0.031	0.389	*rpl2*	0.509 1	0.476 4	0.349 1	0.445	55.13	274	0.138	−0.105	0.354
clpP	0.558 4	0.390 9	0.304 6	0.418	53.12	196	0.167	−0.113	0.342	*rpl20*	0.379 8	0.426 4	0.240 3	0.352	47.31	128	0.124	−0.179	0.312
matK	0.365	0.322 5	0.307 5	0.332	47.32	399	0.161	−0.163	0.317	*rpl22*	0.352 1	0.380 3	0.239 4	0.326	48.42	141	0.197	−0.082	0.386
ndhA	0.436	0.378 7	0.245 2	0.354	44.4	366	0.135	−0.096	0.331	*rpoA*	0.445 1	0.320 5	0.237 4	0.335	48.09	336	0.148	−0.149	0.327
ndhB	0.417 6	0.382 4	0.343 1	0.381	49.94	509	0.165	−0.069	0.363	*rpoB*	0.507 9	0.381 9	0.291 3	0.394	50.32	1 070	0.151	−0.126	0.337
ndhC	0.454 5	0.330 6	0.256 2	0.347	52.18	120	0.185	−0.091	0.333	*rpoC1*	0.495 6	0.378 3	0.266 9	0.381	48.97	681	0.149	−0.139	0.325
ndhD	0.400 8	0.371	0.291 7	0.355	50.04	503	0.138	−0.16	0.306	*rpoC2*	0.428 7	0.357 5	0.307	0.365	50.16	911	0.156	−0.136	0.336
ndhE	0.435 6	0.346 5	0.207 9	0.33	42.47	100	0.147	−0.222	0.268	*rps1*	0.525 2	0.568 3	0.237 4	0.444	47.73	138	0.114	−0.222	0.286
ndhF	0.379 6	0.382 2	0.251	0.338	43.99	784	0.148	−0.148	0.318	*rps12*	0.503 8	0.473 7	0.255 6	0.414	44.37	132	0.122	−0.143	0.313
ndhG	0.418 1	0.339	0.226	0.33	42.63	176	0.148	−0.189	0.266	*rps14*	0.415 8	0.485 1	0.346 5	0.42	46.08	100	0.13	−0.093	0.354
ndhH	0.505 1	0.370 6	0.274 1	0.384	47.57	393	0.158	−0.093	0.346	*rps18*	0.349 1	0.415 1	0.245 3	0.337	38.81	105	0.118	−0.094	0.343
ndhI	0.434 5	0.375	0.244	0.353	41.57	167	0.185	−0.117	0.342	*rps2*	0.447 3	0.434 6	0.27	0.384	51.02	236	0.173	−0.121	0.348
ndhJ	0.509 4	0.389 9	0.377 4	0.426	53.42	158	0.17	−0.165	0.315	*rps3*	0.440 4	0.330 3	0.243 1	0.339	49.61	217	0.159	−0.114	0.352
ndhK	0.421 1	0.443	0.25	0.373	48.56	227	0.154	−0.176	0.312	*rps7*	0.532 1	0.448 7	0.243 6	0.411	44.14	155	0.189	−0.053	0.387
petA	0.512 5	0.35	0.328 1	0.397	56.16	319	0.179	−0.045	0.381	*rps8*	0.414 8	0.4	0.288 9	0.371	43.07	134	0.131	0.044	0.43
petB	0.481 3	0.415 9	0.261 7	0.387	41.33	213	0.191	−0.088	0.347	*ycf1*	0.338 1	0.297 4	0.240 5	0.292	45.58	1 475	0.168	−0.099	0.365
petD	0.208 9	0.512 7	0.392 4	0.39	39.73	146	0.195	−0.035	0.417	*ycf3*	0.497	0.384 6	0.289 9	0.393	51.31	168	0.149	−0.188	0.329
psaA	0.523 3	0.434 1	0.318 2	0.426	49.51	750	0.197	−0.094	0.361	*ycf4*	0.405 4	0.4	0.302 7	0.37	52.06	184	0.154	−0.104	0.349
psaB	0.484 4	0.429 9	0.321 1	0.412	48.23	734	0.179	−0.11	0.353	Average	0.460 5	0.399 1	0.281 6	0.382	47.22	347	0.167	−0.099	0.352

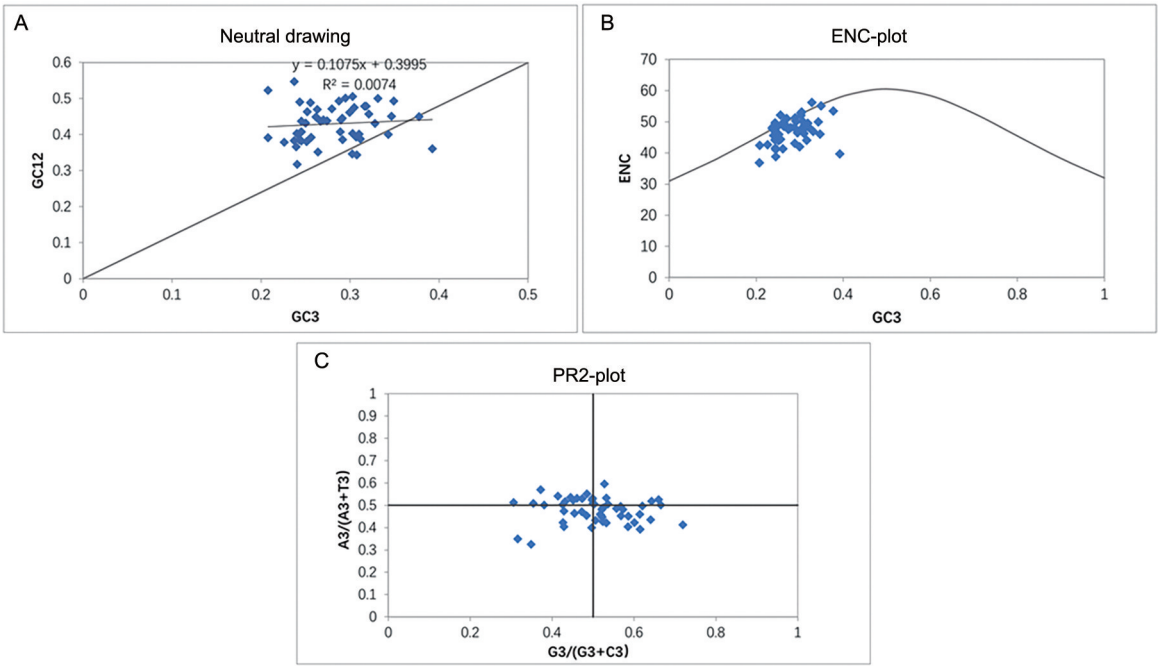

Figure 4　Codon analysis. A: Neutral drawing analysis, the correlation coefficient of GC12 and GC3 is 0.085 8, the correlation is not significant; B: Association analysis of ENC and GC3, most of the genes are below the standard curve and are far from the curve, indicating that the codon preference is greatly affected by natural selection; C: PR2-plot drawing analysis, the distribution of each gene in the four quadrants is uneven, and most of the genes are distributed in the lower and right sides of the plan, indicating that there is inconsistency in the use frequency of the four bases

基使用频率上, 4 个碱基使用频率存在不一致性, 即 T > A, G > C。

3.4　最优密码子分析　根据 51 条 CDS 的 RSCU 值筛选出各库内对应密码子 ΔRSCU > 0.08 的密码子作为高表达密码子, 并将 ΔRSCU > 0.08 且 RSCU > 1 的密码子作为最优密码子[25] (表 4), 得出 RSCU > 1 的高频密码子有 30 个; ΔRSCU ≥ 0.08 的密码子有 24 个, 其中有 3 个以 U 结尾, 有 4 个以 A 结尾, 有 10 个以 C 结尾, 有 7 个以 G 结尾; 最终筛选出最优密码子共有 5 个, 为 CUU、UCU、UCA、CCA、ACU, 其中有 3 个以 U 结尾, 有 2 个以 A 结尾。

4　重复序列分析

红花龙胆叶绿体基因组中共检测到 169 个 SSR (表 5), 其中单核苷酸重复 SSR 最多 (114 个, 67.50%), 其次是二核苷酸 SSR (43 个, 25.44%)、三核苷酸 SSR (3 个, 1.78%)、四核苷酸 SSR (7 个, 4.14%)、六核苷酸 SSR (2 个, 1.18%), 不存在五核苷酸 SSR。在所检测的 SSR 中以 A/T、AT/AT、AAAT/ATTT 和 AATT/AATT 为重复单元的占 81.07%, 表明红花龙胆叶绿体 SSR 偏向 A/T 碱基, SSR 的长度以 8～11 bp 的短序列为主。叶绿体不同区域的 SSR 分布比例分别为 LSC (61.94%)、SSC (18.66%)、IR (19.40%)。有一半的 SSR 位于基因间隔区 (IGS), 58 个 SSR 位于外显子 (exon), 9 个 SSR 位于内含子 (intron)。

5　选择压力分析

Ka/Ks 是评估蛋白质编码基因是否发生适用性进化的有效方法。生物大多数基因的同义核苷酸替换比非同义替换发生得更频繁, 因此 Ka/Ks 值通常小于 1[26]。为进一步研究红花龙胆与龙胆属物种叶绿体基因在进化过程中受到的选择压力, 利用 DnaSP 软件分析红花龙胆、条纹龙胆、粗茎秦艽、条叶龙胆、扁蕾和长春花蛋白编码基因的 Ka/Ks 值 (图 5)。54 个蛋白编码基因在红花龙胆 *vs* 条纹龙胆 (GRvsGS)、红花龙胆 *vs* 粗茎秦艽 (GRvsGC)、红花龙胆 *vs* 条叶龙胆 (GRvsGM)、红花龙胆 *vs* 扁蕾 (GRvsGB) 和红花龙胆 *vs* 长春花 (GRvsCR) 的 Ka/Ks 均值分别为 0.278、0.186、0.176、0.227 和 0.216, 其中绝大多数基因 Ka/Ks 小于 1, 表明龙胆属物种叶绿体基因在长期的进化过程中受到了较强的纯化选择。进一步分析表明, 光合作用相关基因的 Ka/Ks 值除 *psaI* 基因外均小于 1 (图 6); 表达相关基因 *rpl22* 和 *rps11* 在红花龙胆 *vs* 条纹龙胆 Ka/Ks 大于 1, 这些基因在进化过程中受到正向选择作用; 其他功能基因中在各组的比较中 Ka/Ks 均小于 1, 相比于光合作用及表达相关基因, 其他种类的功能基因 Ka/Ks 更小, 受到了相对更强的纯化选择。

将进化趋势较大 (Ka/Ks > 1) 的 *psaI*、*rpl22*、*rps11*

Table 4 Relative synonymous codon usage of each amino acid in *G. rhodantha* chloroplast genome

Amino acid	Codon	RSCU	High expression		Low expression		ΔRSCU	Amino acid	Codon	RSCU	High expression		Low expression		ΔRSCU
			Number	RSCU	Number	RSCU					Number	RSCU	Number	RSCU	
Phe	UUU	1.38	39	1.47	37	1.42	0.05	His	CAU	1.51	18	1.57	10	1.54	0.03
	UUC	0.62	14	0.53	15	0.58	−0.05		CAC	0.49	5	0.43	3	0.46	−0.03
Leu	UUA	2.16	23	1.45	38	2.4	−0.95	Gln	CAA	1.57	31	1.44	25	1.67	−0.23
	UUG	1.1	18	1.14	19	1.2	−0.06		CAG	0.43	12	0.56	5	0.33	0.23
	CUU	1.23	28	1.77	25	1.58	0.19	Asn	AAU	1.52	30	1.4	22	1.63	−0.23
	CUC	0.33	2	0.13	4	0.25	−0.12		AAC	0.48	13	0.6	5	0.37	0.23
	CUA	0.8	12	0.76	9	0.57	0.19	Lys	AAA	1.56	37	1.42	37	1.85	−0.43
	CUG	0.38	12	0.76	0	0	0.76		AAG	0.44	15	0.58	3	0.15	0.43
Ile	AUU	1.52	45	1.23	39	1.72	−0.49	Asp	GAU	1.61	35	1.63	15	1.67	−0.04
	AUC	0.56	25	0.68	5	0.22	0.46		GAC	0.39	8	0.37	3	0.33	0.04
	AUA	0.92	40	1.09	24	1.06	0.03	Glu	GAA	1.53	46	1.44	27	1.64	−0.2
Met	AUG	1	20	1	27	1	0		GAG	0.47	18	0.56	6	0.36	0.2
Val	GUU	1.51	26	1.49	29	1.59	−0.1	Cys	UGU	1.52	6	1.2	4	2	−0.8
	GUC	0.4	10	0.57	7	0.38	0.19		UGC	0.48	4	0.8	0	0	0.8
	GUA	1.59	26	1.49	29	1.59	−0.1	Trp	UGG	1	19	1	16	1	0
	GUG	0.49	8	0.46	8	0.44	0.02	Arg	CGU	1.4	18	1.8	18	2.08	−0.28
Ser	UCU	1.68	21	1.66	11	1.29	0.37		CGC	0.39	4	0.4	1	0.12	0.28
	UCC	0.87	11	0.87	6	0.71	0.16		CGA	1.38	13	1.3	14	1.62	−0.32
	UCA	1.16	15	1.18	7	0.82	0.36		CGG	0.49	2	0.2	3	0.35	−0.15
	UCG	0.58	7	0.55	11	1.29	−0.74		AGA	1.8	16	1.6	14	1.62	−0.02
	AGU	1.31	11	0.87	15	1.76	−0.89		AGG	0.54	7	0.7	2	0.23	0.47
	AGC	0.41	11	0.87	1	0.12	0.75	Gly	GGU	1.3	19	1.1	30	1.82	−0.72
Pro	CCU	1.56	13	1.11	26	1.89	−0.78		GGC	0.43	9	0.52	7	0.42	0.1
	CCC	0.81	14	1.19	10	0.73	0.46		GGA	1.57	23	1.33	22	1.33	0
	CCA	1.1	13	1.11	14	1.02	0.09		GGG	0.7	18	1.04	7	0.42	0.62
	CCG	0.54	7	0.6	5	0.36	0.24	TER	UAA	1.8	1	0.6	2	1.2	−0.6
Thr	ACU	1.58	18	1.8	24	1.71	0.09		UAG	0.6	1	0.6	1	0.6	0
	ACC	0.8	5	0.5	10	0.71	−0.21		UGA	0.6	3	1.8	2	1.2	0.6
	ACA	1.21	12	1.2	16	1.14	0.06	Ala	GCU	1.81	16	1.36	24	1.71	−0.35
	ACG	0.41	5	0.5	6	0.43	0.07		GCC	0.66	9	0.77	4	0.29	0.48
Tyr	UAU	1.63	38	1.73	25	1.72	0.01		GCA	1.12	16	1.36	20	1.43	−0.07
	UAC	0.37	6	0.27	4	0.28	−0.01		GCG	0.42	6	0.51	8	0.57	−0.06

Table 5 SSR in *G. rhodantha* chloroplast genome by MISA

Repeat type	SSR	Number of repeats													Total
		3	4	5	6	7	8	9	10	11	12	13	14	15	
Mononucleotide	A/T	−	−	−	−	−	64	27	5	9	1	1	1	1	109
	C/G	−	−	−	−	−	5								5
Dinucleotide	AG/CT	−	19												19
	AT/AT	−	19	2		3									24
Trinucleotide	AAC/GTT	−	1												1
	AAG/CTT	−	1												1
	AGG/CCT	−	1												1
Tetranucleotide	AAAT/ATTT	2	1												3
	AATT/AATT	1													1
	ACAT/ATGT	2													2
	AGAT/ATCT	1													1
Hexanucleotide	AAGTAC/ACTTGT	2													2

基因从上述12个物种叶绿体基因组注释数据中提取出来, 利用MAFFT比对、MEGA7建树、Figtree可视化后, 得到基于3个基因的系统进化树 (图7), 并通过Geneious可视化3个基因的碱基组成及组内组间碱基变异情况 (图8)。

进一步分析发现, 包括红花龙胆在内的狭蕊组3个物种的*psaI*基因在49 bp处的G碱基突变为T碱基, 导致蛋白中编码该位置的缬氨酸突变为亮氨酸; 在62 bp处的T碱基突变为C, 导致蛋白中编码该位置的异亮氨酸突变为苏氨酸, 这可能会导致狭蕊组物种光合作用的能力发生变化。*rps11*和*rpl22*基因中也有类似的狭蕊组碱基突变现象。狭蕊组的*rpl22*基因在25～34 bp处出现了集中的组间碱基变异, 导致原龙胆属中存在的单核苷酸重复SSR (A) 中断, 并形成了新

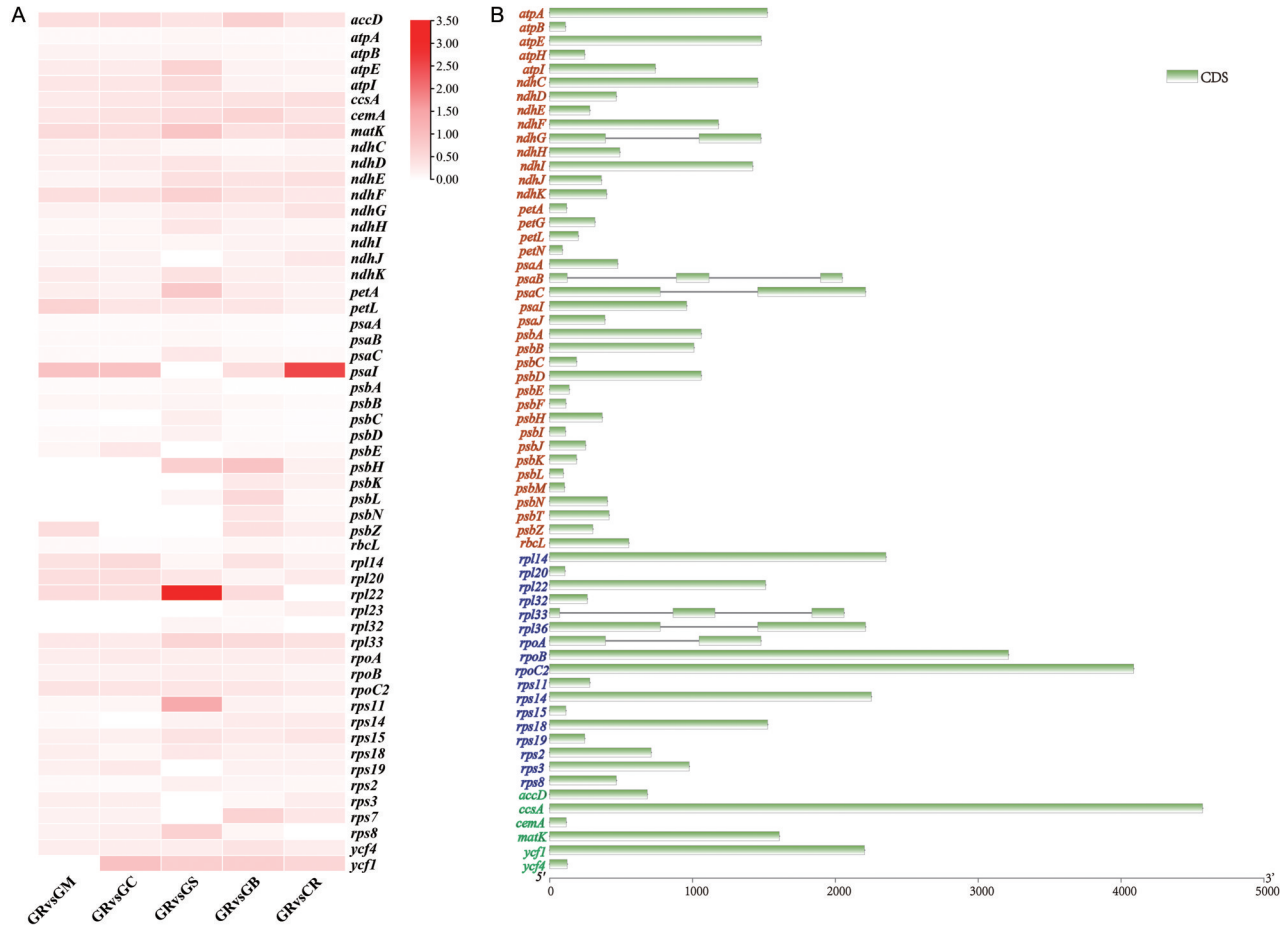

Figure 5　Select pressure analysis of 54 CDS. A: Ka/Ks heatmap of five groups of comparison calculated with DNAsp, Ka/Ks of *psaI*, *rpl22* and *rps11* are greater than 1 and are subject to positive selection. B: CDS structure and length of chloroplast genome

的二核苷酸重复SSR (ATATATATAT)。

讨论

红花龙胆广泛分布于中国西南地区, 是苗族常用药物, 又名青鱼胆草 (苗族药名“锐定谋”), 性味苦寒, 具有清热燥湿、解毒泻火、止咳的功效, 主治湿热黄疸、肺热咳嗽、小便不利, 药效显著, 现已开发了多种与红花龙胆相关的复方配伍剂型。目前关于红花龙胆的研究在功效应用、化学成分、药理活性、质量评价等方面取得了较大进展, 但对其分子生物学水平研究甚少。据文献[27]报道, 该属起源于青藏高原地区, 随着高原隆升和气候变化不断发生物种分化, 并形成分类上一个复杂且较为困难的属。目前关于龙胆属的进化关系问题有很多讨论, 根据《中国植物志》第六十二卷龙胆属记载, 龙胆属下共11组, 包括狭蕊组 (*Sect. Stenogyne*) 和耳褶龙胆组 (*Sect. Otophora* Kusnez.)。Ho等[28]基于形态特征, 将狭蕊组独立成新属—狭蕊龙胆属 (*Metagentiana*), 随后又基于ITS和*trnL* intron序列, 探讨了狭蕊龙胆属与近缘类群的关系, 发现狭蕊龙胆属是多系群, 其与蔓龙胆属 (*Crawfurdia*) 和双蝴蝶属 (*Tripterospermum*) 相互交叉, 三个属均不是独立的属, 共同构成一个单系, 为龙胆属的姊妹群。Favre等[29,30]利用ITS和*atpB-rbcL*序列的分子系统学研究发现, 龙胆属的耳褶龙胆组与蔓龙胆属和双蝴蝶属关系比龙胆属更近, 并将耳褶龙胆组独立为新属耳褶龙胆属 (*Kupferia*), 故耳褶龙胆 *G. otophora* 又名 *K. otophora*。近年来, 随着各项研究深入, 龙胆属物种的进化关系日渐明晰。但龙胆属药用植物在民间药用实践中常相互替代, 不同基原物种间在有效成分、药理作用及功效上仍有明显不同, 且随着市场价值的不断提升, 野生红花龙胆资源急剧锐减, 因此亟待对药材红花龙胆基原植物进行准确鉴定并阐明狭蕊组与龙胆属的进化关系。

密码子偏好性 (codon usage bias)[31-33]对基因的功能和表达有一定影响, 利用叶绿体基因组来完成功能基因的比较分析也成为了当前的难点和重点[34]。相对同义密码子使用度 (RSCU) 也通常被用作密码子偏好性的重要指标, 若RSCU = 1, 则表示密码子使用无偏性, 若RSCU > 1, 则表示该密码子使用频率较高[35], 与

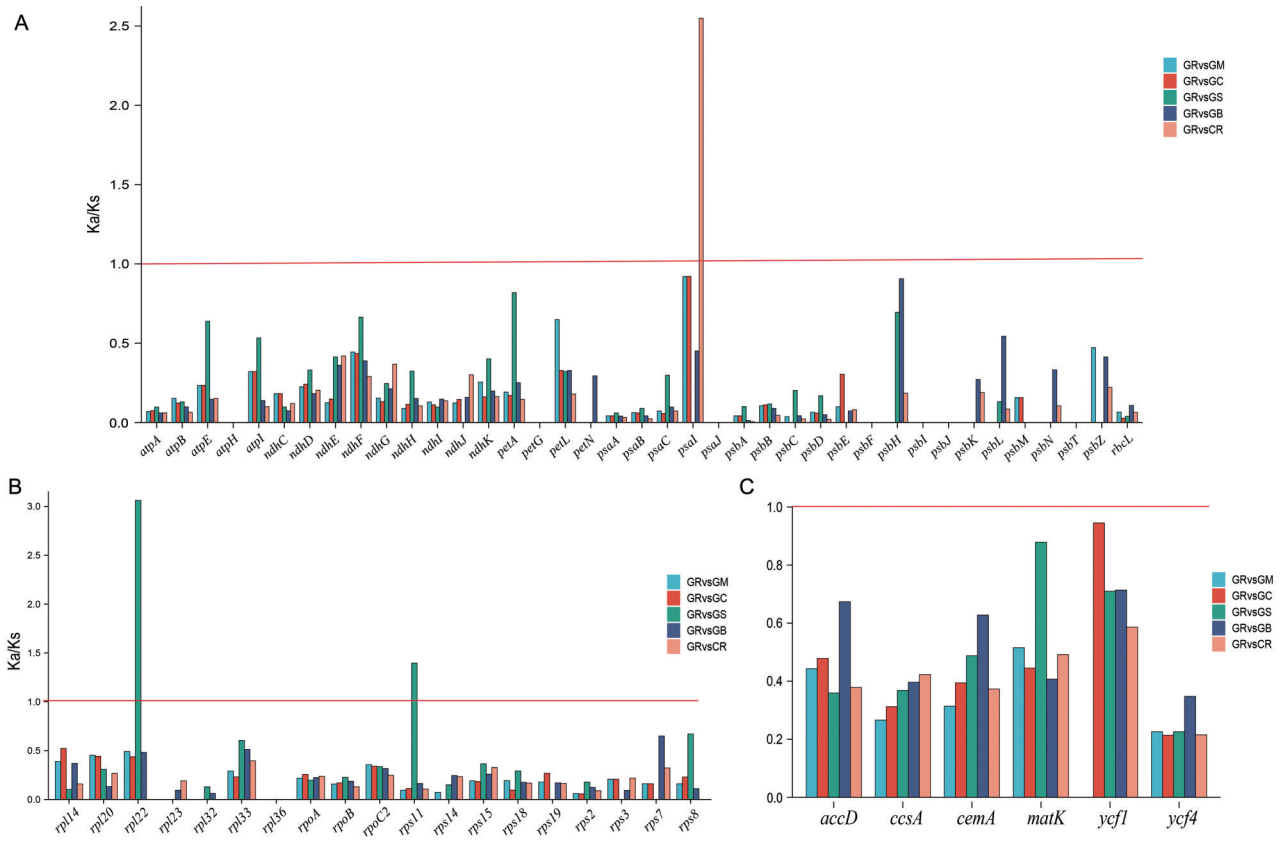

Figure 6 The Ka/Ks ratio analysis. Genes with Ka/Ks greater than 1 were subject to positive selection. A: CDS associated with photosynthesis, the Ka/Ks ratio of *psaI* in group GR *vs* CR was greater than 1; B: CDS associated with expression, the Ka/Ks ratio of *rpl22* and *rps11* in group GR *vs* GS were greater than 1; C: Other CDS

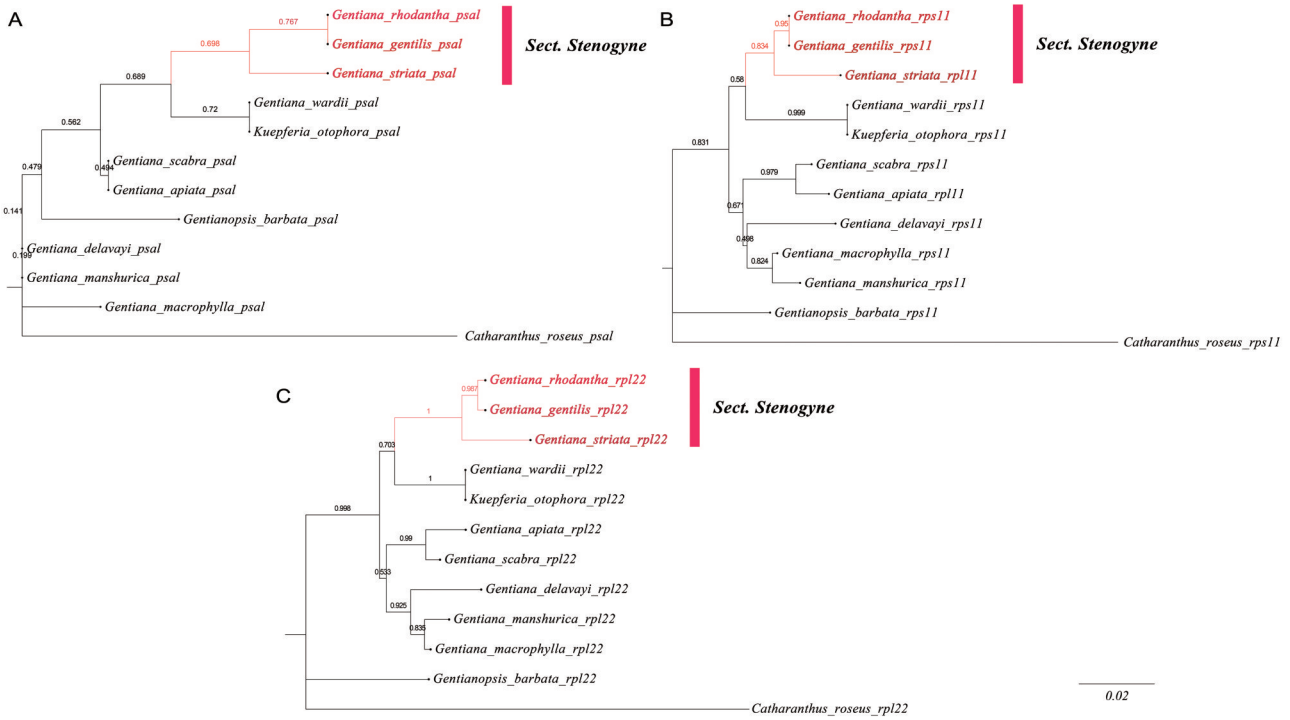

Figure 7 Phylogenetic tree of *psaI* (A), *rps11* (B), *rpl22* (C) based on the NJ method (bootstrap was 1 000 replicates). The three species of *Sect. Stenogyne* are grouped together and can be distinguished from other groups, and basically consistent with the results of the whole genome tree

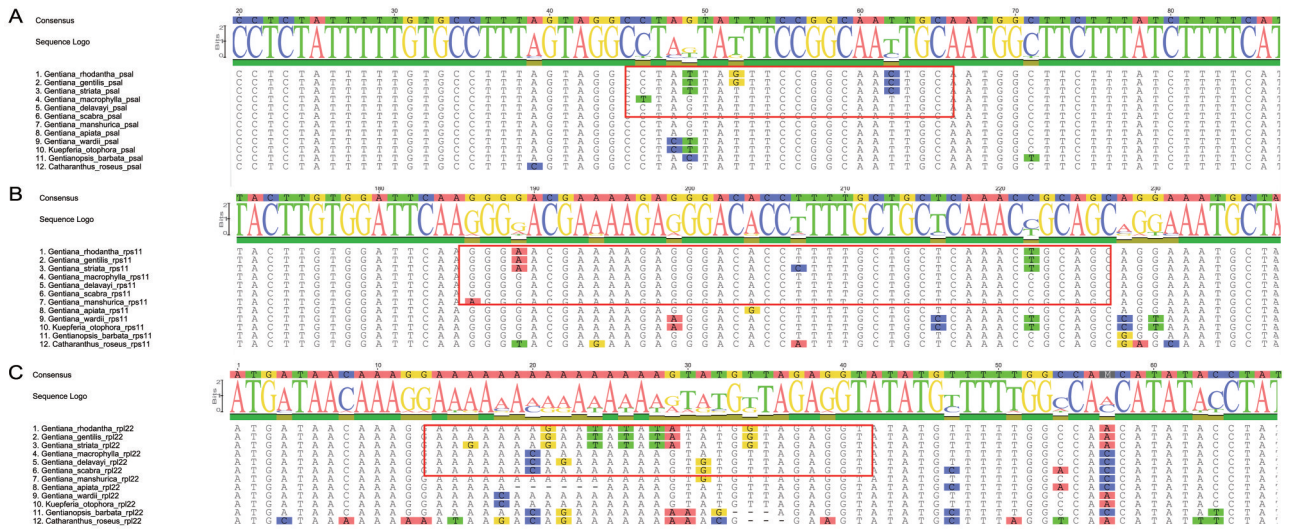

Figure 8　Visualization of base differences in 3 genes using Geneious (A: *psaI*; B: *rps11*; C: *rpl22*). SNP sites unique to *Sect. Stenogyne* are framed in red. *Sect. Stenogyne* always exhibits more unique variation than other species. The size of the bases above indicates the frequency in the genes of 12 species

大部分研究结果相似[36-38], 红花龙胆叶绿体基因组各基因密码子的第3位碱基A和T的使用频率高于G和C, 存在使用偏好。红花龙胆叶绿体基因组密码子偏好性主要受到自然选择的影响, 而定向突变弱于自然选择作用的影响。红花龙胆叶绿体基因组最优密码子以U或A结尾, 与绝大多数高等植物和藻类植物叶绿体基因的最优密码子都以U或A结尾一致[39]。与此同时, 最优密码子及其数量在不同物种间又有所不同, 表明不同物种在进化过程中面临的进化压力可能是由于不同基因特殊的密码子偏好性导致的, 进化压力并不相同。研究表明, 密码子使用偏好性聚类在较小的分类单元中可能提供较为可靠的分类依据, 而当面临大量且复杂的分类单元时, 由于不同基因特殊的密码子偏好性导致这种聚类结果往往不能准确地反映物种亲缘关系[40]。故可推测, 龙胆属系统发育树聚类困难和复杂的部分原因可能是由于不同基因特殊的密码子偏好性。

SSRs也称为微卫星DNA (microsatellites DNA), 是以少数核苷酸(一般1～6个)为基本重复单元构成的简单串联重复序列, 普遍存在于真核生物基因组中, 可作为分子标记运用于群体遗传分析、作物育种等相关研究。红花龙胆叶绿体SSR偏向A/T碱基, 进一步验证了叶绿体基因组序列中的SSRs主要由polyA或polyT所构成, 而较少出现C或G串联重复这一结论[41], SSR的长度以8～11 bp的短序列为主, 且单核苷酸重复的SSR最多, 其中*ycf1*基因上SSR分布最多(10个), 也反映了*ycf1*基因的高变异性。这些SSRs的获得对进一步研究红花龙胆药用植物遗传多样性、群体结构、分子鉴定等方面具有重要意义。

分析适应性进化对于研究基因的结构变化和功能变异有着深远影响。正选择作用可以判断基因是否经历适应性进化, 理解基因的功能和结构的适应关系。54个蛋白编码基因在5组对比中的Ka/Ks均值为0.2左右, 其中绝大多数基因Ka/Ks小于1, 表明龙胆属物种叶绿体基因在长期的进化过程中受到了较强的纯化选择。本研究共检测到3个正选择基因, 与光合作用相关的*pasI*基因在龙胆属红花龙胆*vs*外类群长春花中Ka/Ks大于1, 而在红花龙胆与其他龙胆属植物比较中均小于1, 这可能是与龙胆属植物基本都生长于温带地区的高山地区, 喜温凉气候, 并利用相似的光照条件进行光合作用有关, 而长春花与龙胆属的生长环境、光照条件存在较大差异。与表达相关的*rpl22*和*rps11*在红花龙胆*vs*条纹龙胆中均大于1, 分别编码核糖体大亚基L20和核糖体小亚基S11, 说明这两个基因在红花龙胆在同组的条纹龙胆中有着较大的进化趋势, 深入研究这些基因对讨论红花龙胆的进化关系具有一定的意义。

基于*pasI*、*rps11*、*rpl22*的系统进化树, 可发现3个基因的进化趋势均存在明显的组间差异, 红花龙胆所在的狭蕊组3个物种均聚为一支, 与其他龙胆属的遗传距离相对较远, 且基因序列越长, 聚类结果更接近叶绿体全基因组系统发育树(*pasI*基因111 bp、*rps11*基因417 bp、*rpl22*基因501 bp)。进一步分析发现, 包括红花龙胆在内的狭蕊组3个物种的*pasI*基因在发生了多处的碱基突变, 导致狭蕊组物种*pasI*蛋白发生变化, 这可能会导致狭蕊组物种光合作用的能力发生变化, *rps11*和*rpl22*基因中也有类似的狭蕊组碱基突变现象。此外

狭蕊组中条纹龙胆 (*G. striata*) 表现出相对更多的独立突变, 这与Ho等[28]关于狭蕊龙胆属与近缘类群的讨论契合, 即龙胆属狭蕊组 (*Sect. Stenogyne*) 应独立为新属——狭蕊龙胆属 (*Metagentiana*), 是龙胆属的姊妹群, 且条纹龙胆从中独立出来作为新属 (*Sinogentiana*), 红花龙胆及其他类群仍保留在狭蕊龙胆属中。在传统意义上的龙胆属中, 狭蕊组与其他组的亲缘关系明显更远, 本研究支持了龙胆属中狭蕊组独立成属的观点并提供了新的依据。在*rpl22*基因可发现, 狭蕊组3个物种出现了集中的组间碱基变异, 导致原龙胆属中存在的单核苷酸重复SSR (A) 中断, 并形成了新的二核苷酸重复SSR (ATATATATAT), 针对该SSR片段设计特异性引物, 有望作为鉴别狭蕊组与其他龙胆属物种的候选片段, 为龙胆属复杂的系统发育问题的解决提供新手段。

本研究基于红花龙胆叶绿体基因组测序、组装、注释结果, 通过密码子分析、重复序列分析完成红花龙胆叶绿体基因组的特征分析, 并利用适应性进化分析探讨了复杂的龙胆属分类问题, 结论支持了红花龙胆及狭蕊组物种独立成属的观点, 并筛选出狭蕊组特异性SSR片段。为红花龙胆叶绿体基因工程、遗传多样性分析、分子育种等研究奠定了基础。通过对红花龙胆叶绿体基因组的研究, 激发了研究生科研兴趣、锻炼了动手能力、提升了创新能力, 促进了本草基因组学的教学。

作者贡献: 邓港负责文章撰写及数据分析; 向丽、刘霞负责实验设计及论文修改; 邓港、吴田泽负责数据分析和实验材料的收集; 高冉冉、王梦月指导文章撰写并提出修改意见; 向丽负责论文设计及项目开展。

利益冲突: 所有作者均声明不存在利益冲突。

References

[1] He TN. Gentianaceae in Flora Reipublicae Popularis Sinicae: Vol 62 (中国植物志: 62卷) [M]. Beijing: Science Press, 1988.

[2] Chinese Pharmacopoeia Commission. Pharmacopoeia of the People's Republic of China (中华人民共和国药典) [S]. Beijing: China Medical Science Press, 2015, 1: 151-152.

[3] Wang S, Xie GY, Qin MJ. Advance in pharmaceutical research of *Gentiana rhodantha* [J]. Chin Wild Plant Resour (中国野生植物资源), 2017, 36: 53-59.

[4] Xu M, Wang D, Zhang YJ, et al. Iridoidal glucosides from *Gentiana rhodantha* [J]. J Asian Nat Prod Res, 2008, 10: 491-498.

[5] Yao HQ. Study on Chemical Constituents of *Gentiana rhodantha* (红花龙胆化学成分研究) [D]. Shanghai: Shanghai University of Traditional Chinese Medicine, 2013.

[6] Jiang B. Bai nationality medicinal plant *Gentiana rhodantha* [J]. J Dali Unive (大理学院学报), 2015, 14: 94.

[7] Xu W, Sun AQ, Zhang Z, et al. Morphological traits and resource investigation of *Gentiana rhodantha* in Guizhou province [J]. J Liupanshui Norm Univ (六盘水师范学院学报), 2014, 26: 1-6.

[8] Liu EW, Xu SJ, Xu WF, et al. Analysis of HPLC fingerprint and mangiferin content of *Gentiana rhodantha* from Guizhou [J]. Guizhou Sci (贵州科学), 2020, 38: 40-47.

[9] He Y. A new type of antitussive and expectorant FeiLiKeHeJi [J]. Cent South Pharm (中南药学), 2009, 7: 554-556.

[10] Mu ZQ, Yu Y, Gao H, et al. Research progress on chemical constituents and pharmacological effects of *Sect. Cruciata* [J]. China J Chin Mater Med (中国中药杂志), 2009, 34: 2012-2017.

[11] Shen T, Zhang J, Shen SK, et al. Distribution simulation of *Gentiana rhodantha* in Southwest China and assessment of climate change impact [J]. Chin J Appl Ecol (应用生态学报), 2017, 28: 2499-2508.

[12] Wang B, Gao L, Su YJ, et al. Adaptive evolutionary analysis of chloroplast genes in euphyllophytes based on complete chloroplast genome sequences [J]. Acta Sci Nat Univ Sunyatseni (中山大学学报自然科学版), 2012, 51: 108-113, 146.

[13] Clegg MT, Gaut BS, Learn GH, et al. Rates and patterns of chloroplast DNA evolution [J]. Proc Natl Acad Sci U S A, 1994, 91: 6795-6801.

[14] Ni LH, Zhao ZL, Mi M. Research progress on chloroplast genomes of medicinal plants [J]. J Chin Med Mater (中药材), 2015, 38: 1990-1994.

[15] Kolmogorov M, Bickhart DM, Behsaz B, et al. metaFlye: scalable long-read metagenome assembly using repeat graphs [J]. Nat Methods, 2020, 17: 1103-1110.

[16] Shi L, Chen H, Jiang M, et al. CPGAVAS2, an integrated plastome sequence annotator and analyzer [J]. Nucleic Acids Res, 2019, 47: 65-73.

[17] 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: 1647-1649.

[18] Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability [J]. Mol Biol Evol, 2013, 30: 772-780.

[19] Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets [J]. Mol Biol Evol, 2016, 33: 1870-1874.

[20] Nguyen LT, Schmidt HA, Haeseler AV, et al. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies [J]. Mol Biol Evol, 2015, 32: 268-274.

[21] Beier S, Thiel T, Münch T, et al. MISA-web: a web server for microsatellite prediction [J]. Bioinformatics, 2017, 33: 2583-2585.

[22] Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data [J]. Bioinformatics, 2009, 25: 1451-1452.

[23] Liu XE. A more accurate relationship between 'effective number of codons' and GC3s under assumptions of no selection [J].

Comput Biol Chem, 2013, 42: 35-39.

[24] Sueoka N. Near homogeneity of PR2-bias fingerprints in the human genome and their implications in phylogenetic analyses [J]. J Mol Evol, 2001, 53: 469-476.

[25] Zhang WJ, Zhou J, Li ZF, et al. Comparative analysis of codon usage patterns among mitochondrion, chloroplast and nuclear genes in *Triticum aestivum* L [J]. J Integr Plant Biol, 2007, 49: 246-254.

[26] Makalowski W, Boguski MS. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2 820 orthologous rodent and human sequences [J]. Proc Natl Acad Sci U S A, 1998, 95: 9407-9412.

[27] Ho TN, Liu SW, Lu XF. A phylogenetic analysis of *Gentiana* (Gentianaceae) [J]. Acta Phytotax Sin (植物分类学报), 1996, 34: 505-530.

[28] Ho TN, Chen SL, Liu SW. *Metagentiana*, a new genus of Gentianaceae [J]. Bot Bull Acad Sin, 2002, 43: 83-91.

[29] Favre A, Yuan YM, Küpfer P, et al. Phylogeny of subtribe Gentianinae (Gentianaceae): biogeographic inferences despite limitations in temporal calibration points [J]. Taxon, 2010, 59: 1701-1711.

[30] Favre A, Matuszak S, Sun H, et al. Two new genera of Gentianinae (Gentianaceae): *Sinogentiana* and *Kuepferia* supported by molecular phylogenetic evidence [J]. Taxon, 2014, 63: 342-354.

[31] Romero H, Zavala A, Musto H. Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern of selective forces [J]. Nucleic Acids Res, 2000, 28: 2084-2090.

[32] Nie XJ, Deng PC, Feng KW, et al. Comparative analysis of codon usage patterns in chloroplast genomes of the Asteraceae family [J]. Plant Mol Biol Rep, 2014, 32: 828-840.

[33] Shen ZN, Gan ZM, Zhang F, et al. Analysis of codon usage patterns in citrus based on coding sequence data [J]. BMC Genomics, 2020, 21: 234.

[34] Li JF, Li XY, Wang Y, et al. Analysis on codon usage bias of chloroplast genome in *Catalpa fargesii* Bur. f. *duclouxii* [J]. Genomics Appl Biol (基因组学与应用生物学), 2021. https://kns.cnki.net//kcms/detail/45.1369.Q.20211103.1431.002.html.

[35] Zhao S, Deng LH, Chen F. Codon usage bias of chloroplast genome in *Kandelia obovata* [J]. J Forest Environ (森林与环境学报), 2020, 40: 534-541.

[36] Yang GF, Su KL, Zhao YR, et al. Analysis of codon usage in the chloroplast genome of *Medicago truncatula* [J]. Acta Pratac Sin (草业学报), 2015, 24: 171-179.

[37] Ye YJ, Ni ZX, Bai TD, et al. The analysis of chloroplast genome codon usage bais in *Pinus massoniana* [J]. Genomics Appl Biol (基因组学与应用生物学), 2018, 37: 4464-4471.

[38] Li DM, Lv FB, Zhu GF, et al. Analysis on codon usage of chloroplast genome of *Oncidium Gower* Ramsey [J]. Guangdong Agric Sci (广东农业科学), 2012, 39: 61-65.

[39] Hu XY, Xu YQ, Han YZ, et al. Codon usage bias analysis of the chloroplast genome of *Ziziphus jujuba* var. *spinosa* [J]. J Forest Environ (森林与环境学报), 2019, 39: 621-628.

[40] Christianson ML. Codon usage patterns distort phylogenies from or of DNA sequences [J]. Am J Bot, 2005, 92: 1221-1233.

[41] Kuang DY, Wu H, Wang YL, et al. Complete chloroplast genome sequence of *Magnolia kwangsiensis* (Magnoliaceae): implication for DNA barcoding and population genetics [J]. Genome, 2011, 54: 663-673.