

薯蓣属叶绿体基因组特征、密码子偏好性与机器学习分子鉴别研究

郑梦迪¹, 徐笑雨¹, 黄美琳², 高明晨¹, 赵 阳¹, 宋 茜¹

1. 西安医学院药学院, 陕西 西安 710021

2. 抚州市临川区市场监督管理局, 江西 抚州 344400

摘要: **目的** 结合叶绿体基因组、密码子偏好性与 DNA 条形码、机器学习技术, 为薯蓣属 *Dioscorea* 系统发育、遗传多样性及物种鉴定提供理论指导。**方法** 通过 IRscope 比对 11 种植物的叶绿体基因结构, 并构建系统发育树; 用 CodonW、CUSP 等工具分析密码子偏好性; 收集薯蓣材料并扩增测序 DNA 条形码序列, 基于成熟酶 K 基因 (maturase K gene, *matK*)、光系统 II 蛋白 D1 基因-组氨酸 tRNA 基因间隔区 (photosystem b a protein gene-transfer RNA-histidine intergenic spacer, *psbA-trnH*)、核酮糖-1,5-二磷酸羧化酶/加氧酶大亚基基因 (ribulose-bisphosphate carboxylase/oxygenase large subunit gene, *rbcL*) 条形码和机器学习算法进行分子鉴别。**结果** 薯蓣属叶绿体基因组保守稳定, 密码子显著 A/T 偏好, 第 3 位倾向 A/U 结尾。自然选择是影响密码子偏差的主因, 共有最优密码子为 GGA 和 UCA。3 种条形码均能成功鉴别物种, 基于 BLOG 算法的单一条形码鉴别成功率 100%, WEKA 的 SMO 和 NaïveBayes 分类器鉴别率较高。**结论** 薯蓣叶绿体基因组保守, 密码子偏好模式以自然选择为主, 为薯蓣属基因表达调控、物种鉴定、资源保护等提供依据和指导。

关键词: 薯蓣属; 叶绿体全基因组; 密码子偏好性; DNA 条形码; 机器学习; 分子鉴别

中图分类号: R286.12 文献标志码: A 文章编号: 0253-2670(2026)02-0640-12

DOI: 10.7501/j.issn.0253-2670.2026.02.023

Research on chloroplast genomic characteristics, codon usage bias, and machine learning for molecular identification in *Dioscorea*

ZHANG Mengdi¹, XU Xiaoyu¹, HUANG Meilin², GAO Mingchen¹, ZHAO Yang¹, SONG Xi¹

1. School of Pharmacy, Xi'an Medical University, Xi'an 710021, China

2. Linchuan District Market Supervision Bureau, Fuzhou 344400, China

Abstract: Objective To provide theoretical guidance for the phylogenetics, genetic diversity, and species identification of *Dioscorea* by integrating chloroplast genomics, codon usage bias, DNA barcoding, and machine learning technologies. **Methods** The chloroplast gene structures of 11 plant species were compared via IRscope and a phylogenetic tree was constructed. Codon usage bias was analyzed using tools such as CodonW and CUSP. *Dioscorea* materials were collected, and DNA barcode sequences were amplified and sequenced. Molecular identification was performed based on the maturase K gene (*matK*), photosystem b a protein gene-transfer RNA-Histidine intergenic spacer (*psbA-trnH*), and ribulose-bisphosphate carboxylase/oxygenase large subunit gene (*rbcL*) barcodes using machine learning algorithms. **Results** The chloroplast genomes of *Dioscorea* were found to be conserved and stable. Codon usage showed a significant A/T bias, with the third codon position favoring A/U endings. Natural selection was the primary factor influencing codon bias, and the common optimal codons identified were GGA and UCA. All three barcodes successfully discriminated species. Single-barcode identification using the BLOG algorithm achieved 100% success rate, while the SMO and NaïveBayes classifiers in WEKA demonstrated high identification accuracy. **Conclusion** The chloroplast genome of *Dioscorea* is conserved, and natural selection dominates its codon usage pattern. This study provides a basis and guidance for research on gene expression regulation, species identification, and resource conservation of *Dioscorea*.

Key words: *Dioscorea* L.; chloroplast genome; codon usage bias; DNA barcoding; machine learning; molecular identification

收稿日期: 2025-10-02

基金项目: 陕西省自然科学基金基础研究计划项目 (2021JQ-782); 西安医学院 2024 年度科研能力提升计划项目 (2024NLTS122); 西安医学院 2022 年度科研能力提升计划项目 (2022NLTS084)

作者简介: 郑梦迪 (1988—), 女, 博士, 副教授, 主要从事生物资源鉴定与评价方向研究。E-mail: 670677492@qq.com

薯蓣属 *Dioscorea* L. 隶属薯蓣科, 是重要的药食同源植物, 以根或根茎入药或食用, 《中国药典》2020 版收录其 8 个物种^[1]。该属富含薯蓣皂苷元, 为甾体激素药物关键原料, 具有极高药用价值。常见药用种包括薯蓣 *D. polystachya* Turcz. (山药)、穿龙薯蓣 *D. nipponica* Makino (穿山龙)、盾叶薯蓣 *D. zingiberensis* C. H. Wright (黄姜) 等。此外, 需注意该属植物黄独 *D. bulbifera* L. 具毒性^[1]。中药材山药本草考证表明“薯蓣”为其正名, 历代文献记载如《山海经》中“藟蕒”^[2]、《植物名实图考》中“脚板薯”实为参薯 *D. alata* L.^[3]、《五杂俎》对药用与食用山药的分类^[4]均指向薯蓣。现代资源调查发现, 商品“山药”还包含山薯 *D. fordii* Prain et Burkill、参薯、日本薯蓣 *D. japonica* Thunb. 及褐苞薯蓣 *D. persimilis* Prain et Burkill 等未入《中国药典》物种^[5]。

叶绿体基因组因其结构稳定、长度适中及进化速率合理等特点, 在物种起源、系统进化及比较基因组学研究中具有显著优势^[6]。叶绿体基因组的反向重复区边界 (inverted repeat boundary, IR)、简单序列重复位点 (simple sequence repeat locus, SSR) 及变异位点等信息可用于解析物种的遗传多样性、群体结构及亲缘关系, 对植物种质资源保护具有重要价值^[7]。密码子偏好性源于密码子简并性, 是叶绿体基因组进化中的重要机制, 直接影响基因表达水平、mRNA 稳定性、翻译效率及蛋白质折叠过程。其演化遵循“选择-突变-遗传漂变”理论, 深入研究有助于揭示基因表达调控与进化动力^[8]。

DNA 条形码技术利用标准化的短基因片段, 凭借其变异位点丰富、扩增简便、鉴定准确等优势, 在物种鉴别中广泛应用。其核心在于标准化、极简性和可扩展性, 显著弥补了传统形态分类学的局限性, 尤其在中药资源鉴定与质量标准化方面展现出广阔前景^[9]。近年来, 机器学习为生物分类与遗传鉴定提供了创新方法, 能够高效挖掘 DNA 序列中的分类特征, 精准识别关键变异位点, 提升物种鉴别精度。本研究基于 BLOG 算法与 WEKA 平台, 系统评估不同分类器在物种鉴别中的性能^[10-11]。将机器学习与 DNA 条形码技术相结合, 可进一步提高对近缘物种的鉴别效率与可靠性。

目前, 薯蓣属植物的遗传背景研究仍较为薄弱, 叶绿体基因组信息不足, 限制了该属资源的系统开发与利用^[12]。为此, 本研究拟系统解析薯蓣属植物的叶绿体基因组结构特征与密码子使用偏好

性; 整合常用 DNA 条形码成熟酶 K 基因 (maturase K gene, *matK*)、光系统 II 蛋白 D1 基因—组氨酸 tRNA 基因间隔区 (photosystem b a protein gene-transfer RNA-histidine intergenic spacer, *psbA-trnH*)、核酮糖-1,5-二磷酸羧化酶/加氧酶大亚基基因 (ribulose-bisphosphate carboxylase/oxygenase large subunit gene, *rbcL*), 并引入机器学习方法进行物种鉴别, 进一步与传统遗传距离 (Kimura 2-parameter model, 木村双参数模型) 及邻接 (neighbor-joining, NJ) 法构建结果进行比较, 以填补该属在叶绿体基因组进化、密码子使用模式及智能鉴别方面的研究空白, 为薯蓣属物种鉴定、资源保护与药用安全提供科学依据。

1 材料与仪器

1.1 材料

从 GenBank 获取 11 种薯蓣属植物薯蓣 (NC037716.1)、参薯 (OP787125.1)、穿龙薯蓣 (NC057065.1)、黄独 (PP130732.1)、日本薯蓣 (NC057066.1)、盾叶薯蓣 (NC027090.1)、粉背薯蓣 (NC037717.1)、白薯蓣 (NC066182.1)、褐苞薯蓣 (NC057257.1)、福州薯蓣 (OQ525994.1)、山薯 (NC070252.1) 的完整叶绿体基因组序列。使用 CPGAVAS2 注释, 得到各物种 84~86 条 CDS 序列。进一步筛选: CDS 序列唯一、长度 ≥ 300 bp、具有标准起始密码子 (ATG) 和终止密码子 (TAA/TGA/TAG) 且内部无终止密码子。最终每个物种保留 53 条 CDS 用于密码子偏好性分析。另从 GenBank 获取 34 条 *rbcL* (表 1)、42 条 *matK* (表 2) 和 30 条 *psbA-trnH* (表 3) 序列 (含 1 份江西庐山薯蓣样本)。所有 fasta 格式序列经 clustal 比对并筛选后保留。

1.2 仪器

S1000 PCR 仪 (Bio-Rad 公司, 美国)、KH-300B 台式冷冻离心机 (昆山禾创超声仪器有限公司, 中国)、微量移液器 (Eppendorf 公司, 德国)。

2 方法

2.1 DNA 条形码获取与测序

取 100 mg 新鲜叶片, 超纯水清洗, 75% 乙醇消毒后液氮冷冻研磨。使用天根生化科技 DNA 提取试剂盒提取 DNA, -20°C 保存。通用引物对样本的 *rbcL*、*matK*、*psbA-trnH* 序列进行 PCR 扩增^[13]。PCR 产物由西安擎科泽西生物科技公司完成正向测序。

表 1 *rbcL* 条形码分析样本信息Table 1 Sample information of *rbcL* barcode analysis

物种	拉丁名	来源	登录号/样本号
薯蓣	<i>D. polystachya</i>	GenBank 江西庐山	JQ260201、JQ260202、JQ260204、JQ260205 JX250210
参薯	<i>D. alata</i>	GenBank	OM981345、OM981346、OM981347、OM981344
穿龙薯蓣	<i>D. nipponica</i>	GenBank	HQ637729、HQ637733、HQ637734
黄独	<i>D. bulbifera</i>	GenBank	JQ260176、JQ260177、JQ260178、JQ260179、JQ260180
日本薯蓣	<i>D. japonica</i>	GenBank	KR072426、KR072427、KR072428、KR072429
盾叶薯蓣	<i>D. zingiberensis</i>	GenBank	JQ260123、JQ260124、JQ260125、JQ260126、JQ260127
白薯茛	<i>D. hispida</i>	GenBank	KU865501、KU865504、KU865505、MG707302
山薯	<i>D. fordii</i>	GenBank	JQ260227、JQ260229、JQ260230、JX139771

表 2 *matK* 条形码分析样本信息Table 2 Sample information of *matK* barcode analysis

物种	拉丁名	来源	登录号/样本号
薯蓣	<i>D. polystachya</i>	GenBank 江西庐山	JQ260051、JQ260053、JQ260054、JQ260056 JX250211
参薯	<i>D. alata</i>	GenBank	JQ260095、JQ260096、JQ260097、JQ260098、JQ260099
穿龙薯蓣	<i>D. nipponica</i>	GenBank	JQ259957、HQ637579、JQ259959、JQ259960、HQ637580
黄独	<i>D. bulbifera</i>	GenBank	JQ260026、JQ260027、JQ260030、JQ260031、JQ260032
日本薯蓣	<i>D. japonica</i>	GenBank	KR072296、KR072297、KR072298
盾叶薯蓣	<i>D. zingiberensis</i>	GenBank	JQ259972、JQ259974、JQ259976、JQ259978、JQ259979
白薯茛	<i>D. hispida</i>	GenBank	JQ260046、MG712586、HQ637666、JQ733672、JQ733718
褐苞薯蓣	<i>D. persimilis</i>	GenBank	JX501470、JX501471、JX501472、JX501479、JX501480
福州薯蓣	<i>D. futschauensis</i>	GenBank	JQ733671、DQ974175、HQ637624、JQ260004

表 3 *psbA-trnH* 条形码分析样本信息Table 3 Sample information of *psbA-trnH* barcode analysis

物种	拉丁名	来源	登录号/样本号
薯蓣	<i>D. polystachya</i>	GenBank 江西庐山	HQ637973、HQ637974、HQ637975 JX250212
参薯	<i>D. alata</i>	GenBank	HQ638018、KR072383、KR072382、GQ265165、HQ638016
穿龙薯蓣	<i>D. nipponica</i>	GenBank	JQ260257、JQ260258、JQ260259
黄独	<i>D. bulbifera</i>	GenBank	KR072387、KR072388、KR072389
日本薯蓣	<i>D. japonica</i>	GenBank	KR072296、KR072297、KR072298
盾叶薯蓣	<i>D. zingiberensis</i>	GenBank	HQ637890、HQ637893、HQ637894、HQ637889、GQ265154
粉背薯蓣	<i>D. colletii</i>	GenBank	HQ637915、HQ637916、HQ637917
白薯茛	<i>D. hispida</i>	GenBank	MZ396292、GQ265145、HQ637964

2.2 叶绿体全基因组的比较分析

使用 IRscope 在线软件分析 11 种薯蓣属植物注释后的叶绿体基因组，比较其小单拷贝区 (small single copy region, SSC)、大单拷贝区 (large single copy region, LSC) 和反向重复区 (inverted repeat region, IR) 的边界差异，评估基因组边界稳定性^[14]。利用 MISA 在线网站分析 11 种植物叶绿体基因组的 SSR 分布。运用 mVISTA 在线工具 (Shuffle-LAGAN 模式) 进行序列比对分析^[15]。采用

MEGA11.0 软件对 11 种植物的叶绿体全基因组序列进行比对和校正，并构建 NJ 系统发育树。

2.3 密码子偏好性分析

使用 CodonW 和 mEMBOSS 计算密码子参数： $A_3/T_3/C_3/G_3$ 、 GC_{3s} 、 $GC_1/GC_2/GC_3$ 、 GC_{all} 、CBI、Fop、CAI、RSCU 及 $ENC^{[16]}$ ，通过 ChiPlot 工具分析组内相关性及其显著性。以 GC_3 为横坐标、 GC_{12} 为纵坐标进行中性绘图分析：若基因沿斜率 ≈ 1 的对角线分布，表明突变压力主导密码子偏好；否则自然选择

影响更大^[17]；以 ENC 为纵坐标、GC_{3s} 为横坐标（参照标准曲线 $ENC = 2 + GC_3^2 + 29/[GC_3^2 + (1 - GC_3)^2]$ ）进行 ENC-plot 分析，若点位于曲线上方或附近，突变压力为主，反之自然选择主导^[18]；以 $A_3/(A_3 + T_3)$ 和 $G_3/(G_3 + C_3)$ 为坐标进行 PR2-plot 分析：中心点（A=T，C=G）反映无偏好性，分布偏离中心则受自然选择等多因素影响^[19]。运用 CodonW 对 59 个信息密码子（排除 UGG/AUG/终止密码子）的 RSCU 进行对应分析，提取变异率递减的正交轴。按 ENC 值选取高/低表达基因（两端各 10%），计算 $\Delta RSCU$ （高低库差值）。若密码子 $\Delta RSCU > 0.08$ 且 $RSCU > 1$ ，则定义为最优密码子^[20]。

2.4 基于 DNA 条形码及机器学习的分子鉴别

利用 MEGA 11.0 对薯蓣属 11 种植物的 3 个 DNA 条形码（*rbcL*、*matK*、*psbA-trnH*）进行序列比对与分析，计算种内序列的 GC 含量、长度、保守位点（C）、变异位点（V）及简约信息位点（PI），以及种间序列的单态位点（S）、C、V 和 PI。基于 Kimura 双参数模型计算遗传距离，采用 NJ 法构建

系统发育树，并通过 bootstrap 1000 次检验分支支持率。利用 BLOG 分析序列，提取植物识别规则并制表。应用 WEKA 平台的 4 种分类器（Naïve Bayes、SMO、JRip、J48）进行学习分析，计算物种鉴别成功率。

3 结果与分析

3.1 叶绿体全基因组序列比较分析

3.1.1 简单重复序列 SSR 分析 如图 1 所示，在叶绿体基因组中，识别出符合条件的 SSR 位点总数范围为 37~51，检测到 3 种类型的 SSR（单核苷酸、二核苷酸和三核苷酸序列），各物种通过 SSR 的数量和类型相互区分。盾叶薯蓣的 SSR 位点数最多为 51，白薯蓣的 SSR 位点数最少为 37。薯蓣和日本薯蓣都有 4 个二核苷酸，单核苷酸重复出现的频率高于其他类型的核苷酸，大部分由 A/T 组成，占比 82.22%~89.66%，仅黄独和盾叶薯蓣存在有特殊的单核苷酸 G/C。这反映了叶绿体基因组的 A/T 碱基偏好性，导致多数 SSR 位点由 polyA 或 polyT 类型的重复序列构成。

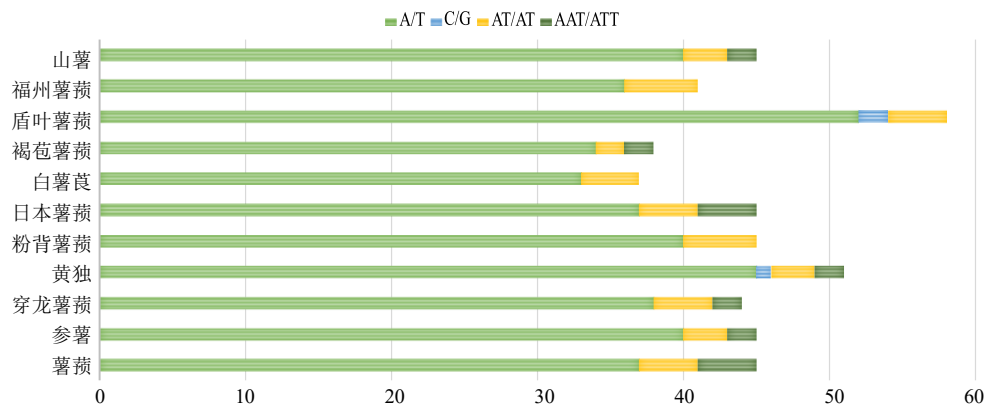


图 1 11 种薯蓣属植物的 SSR 分析

Fig. 1 SSR analysis of 11 species of *Dioscorea* species

3.1.2 IR 边界收缩及扩张分析 通过在线网站 IRscope 比较分析了薯蓣属 11 个物种的叶绿体基因组 LSC、SSC 和 IR 区域的边界区域收缩与扩张，结果（图 2）发现叶绿体基因组的长度差异较小，分布范围为 152 963~153 970 bp，序列整体上相似度较高，边界附近的基因类型相同，相对保守，但在边界区有收缩和扩张现象，物种之间具有一定差异。11 个物种的 IRb/SSC 边界都位于 *ycf1* 基因内部。盾叶薯蓣、白薯蓣、福州薯蓣及黄独的 SSC/IRa 边界距离 *ndhF* 基因 85、25、6、4 bp，且 *ndhF* 基因完全包含于 SSC 中。其余植物

的 *ndhF* 基因大部分位于 SSC 区略向 IRa 区扩张。IRa/SSC 边界均位于 *trnH* 和 *psbA* 之间，但基因距离边界长度有差异。

3.1.3 叶绿体全基因组变异分析 以薯蓣为参考，对比分析 11 种薯蓣属植物叶绿体全基因组。图 3 显示多数基因组区域相似度高，但部分区域有差异，SSC 区变异率最高，因功能基因少，选择压力低；LSC 区功能基因多，保守性强；IR 区基因功能稳定，与光合作用和能量代谢相关。所以差异程度为 $SSC > LSC > IR$ 。非编码区序列变异大于编码区，编码区总体保守，但 *ycf1*、*ycf2*、*psba* 等基因编码区变异较大。

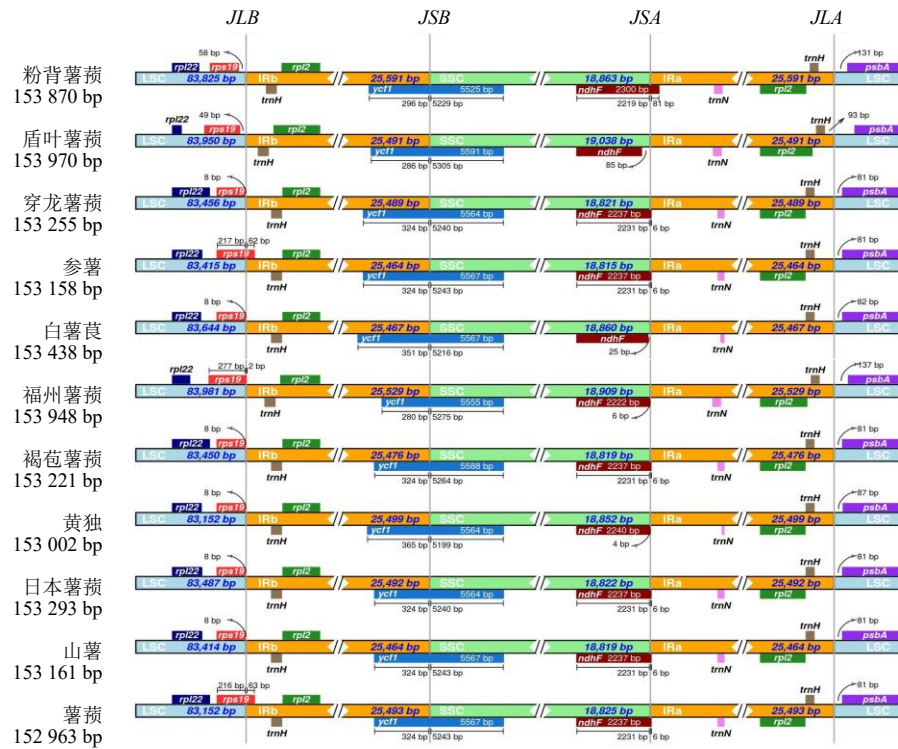
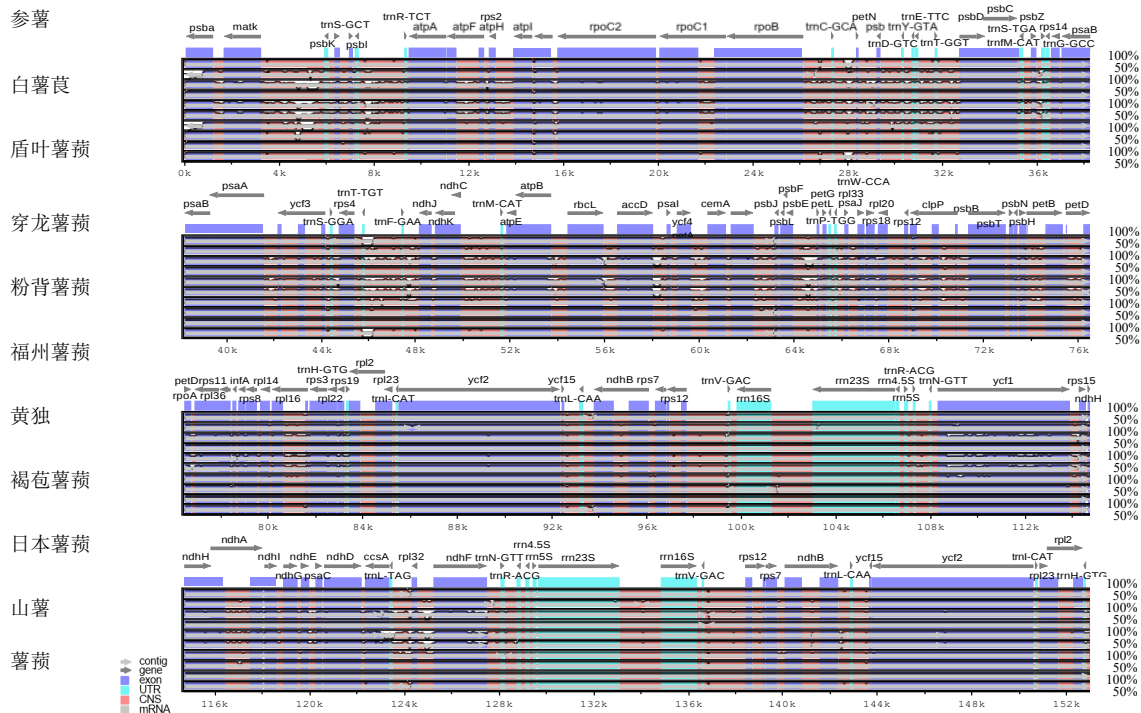


图2 11种薯蓣属植物叶绿体基因组LSC、SSC和IR边界区域比较

Fig. 2 Comparison of LSC, SSC, and IR boundary regions in chloroplast genomes of 11 *Dioscorea* species

基因方向用灰色箭头表示。

Gene orientations are denoted by gray arrows.

图3 mVISTA对薯蓣属植物叶绿体基因组进行可视化表示

Fig. 3 Visual representation of *Dioscorea* species chloroplast genome utilizing mVISTA

3.1.4 系统发育分析 如图 4 所示, 盾叶薯蓣、粉背薯蓣及福州薯蓣首先与其他 8 个物种区分开, 形成独立分支。其次, 参薯和山薯聚为一小分支, 薯蓣和日本薯蓣共处一小分支, 可见, 相较其他物种, 这 2 个小分支上物种彼此间亲缘关系更为接近。

3.2 密码子偏好性分析

3.2.1 密码子相关偏性指数分析 11 种薯蓣属植物叶绿体基因组均有 53 条 CDS。GC 含量为 37.75%~37.87%, GC_1 (46.13%~46.35%) $> GC_2$ (39.13%~39.38%) $> GC_3$ (27.73%~28.09%), 各位置 GC 含量均小于 50.00%, 表明密码子构成偏向 A/T, 第 3 位倾向 A/U 结尾。ENC 平均值在 48.40~48.49, 均大于 35, 说明密码子使用无显著偏好。

3.2.2 相关性分析 由图 5 分析表明, GC 含量与 GC_1 、 GC_2 、 GC_3 呈极显著正相关 (0.27~0.85), 但与 ENC 无关; GC_1 与 GC_2 显著相关 (0.45~0.48), 二者均与 GC_3 无关, 提示密码子第 1、2 位碱基组成保守, 第 3 位变异较大。所有物种中 CAI 与 CBI、Fop 呈正相关, GC_{3s} 与 ENC 呈极显著正相关, 说明密码子偏好性受 GC 含量影响, 高 GC

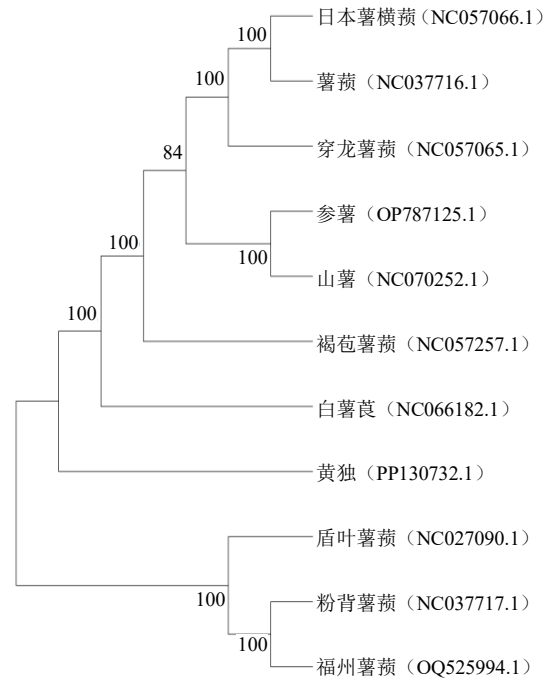
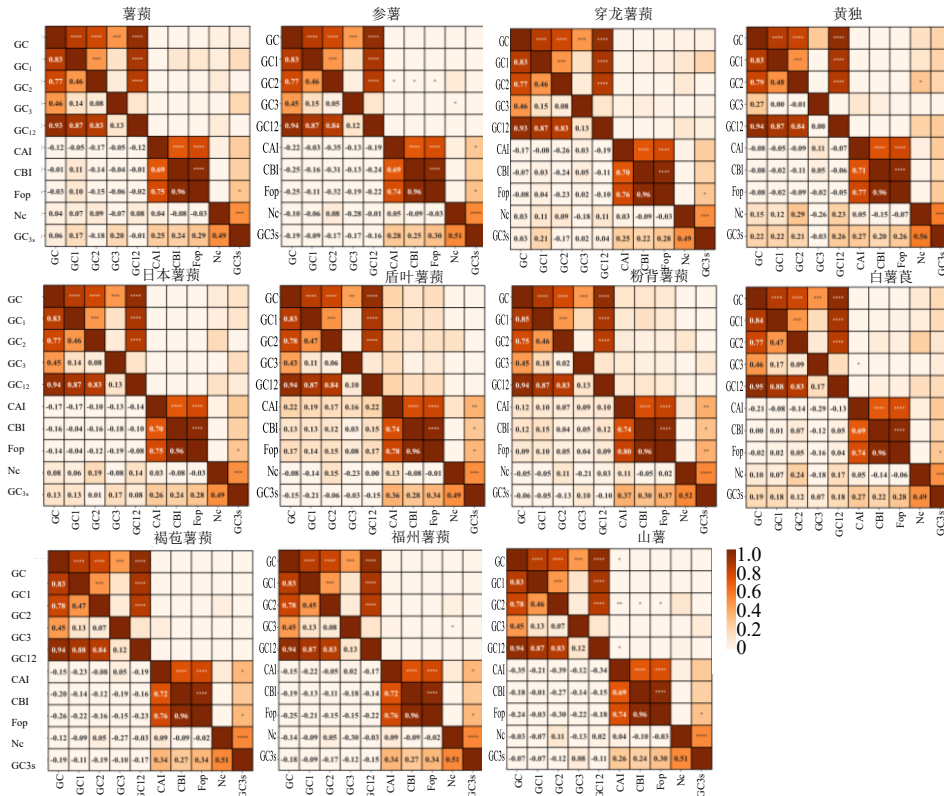


图 4 薯蓣属叶绿体全基因组系统发育分析

Fig. 4 Phylogenetic analysis of chloroplast genome in *Dioscorea* species



* $P < 0.05$ ** $P < 0.01$ *** $P < 0.001$.

图 5 薯蓣属植物叶绿体基因组密码子使用指标的相关分析热图

Fig. 5 Correlation analysis heatmap of codon usage indicators in *Dioscorea* species chloroplast genomes

区域倾向使用特定密码子以优化表达。此外,山薯和参薯的 GC_2 与 CAI 相关,白薯茛的 GC_2 与 ENC 相关,反映不同物种因进化适应存在 GC 含量差异。

3.2.3 中性绘图分析 如图 6 所示,11 物种基因分布均偏离对角线(斜率 0.111 7~0.188 7),表明突变对密码子偏好性影响微弱(11.17%~18.87%);而 R^2 值极低(0.006 5~0.019 4),揭示自然选择压力作用更显著(0.65%~1.94%)。 GC_{12} 与 GC_3 相关性弱,证实密码子 3 个位点碱基组成存在显著差异。综上,薯茛属密码子偏好性主要受自然选择主导,各物种受影响程度排序为日本薯茛>白薯茛>薯茛>粉背薯茛>穿龙薯茛>福州薯茛>参薯>褐苞薯茛>山薯>黄独>盾叶薯茛。

3.2.4 ENC-plot 绘图分析 如图 7 所示,ENC 和 GC_3 在 11 种薯茛属药用植物叶绿体基因组中分布相似,大部分基因点 ENC 值>30,集中在曲线两侧,多数位于曲线下方,表明突变压力对密码子偏好影响小,主要受自然选择影响。38%~42%的基因数靠近标准曲线,58%~62%远离,进一步证明自然选择主导密码子偏好,与中性图分析一致。

3.2.5 PR2-plot 绘图分析 如图 8 所示,11 个叶绿体基因组密码子第 3 位碱基分布不均,多数位于下半部分(第 3、4 象限),表明 A/T 碱基更倾向 T;左右分

布中,多数位于右半部分(第 1、4 象限),显示 G/C 碱基更倾向 G。综合分析表明密码子第 3 位碱基倾向 G/T,自然选择是密码子偏好的主导因素。

3.2.6 对应性分析 由图 9 表明,部分密码子靠近基因轴,显示突变压力影响叶绿体基因密码子偏好。轴 1 为主要变异源,轴顺序增加时主因子轴占比递减,说明密码子偏好由多因子共同作用。日本薯茛和福州薯茛的 RSCU 值在 Axi1 和 Axi2 上分布接近,密码子使用模式相似;参薯、盾叶薯茛等的 RSCU 值分布集中,表明密码子使用高度保守。

3.2.7 最优密码子的确定及 RSCU 分析 本研究基于 11 种薯茛属植物叶绿体基因组的 RSCU 和 ENC,筛选出具有高表达和高频率特征的最优密码子。结果表明其最优密码子数为 5~12 个,均以 U/A 结尾。其中薯茛最多,福州薯茛最少,穿龙薯茛、参薯、山薯各有 9 个。黄独和粉背薯茛分别有独特 GAU 和 UUU 最优密码子。共享最优密码子为 GGA 和 UCA。RSCU 是衡量密码子偏好性的重要指标。结果显示,11 个叶绿体基因组的 RSCU 值在 0.26~1.93,有 30 个高频密码子,其中 29 个以 U/A 结尾(96.67%),仅 1 个以 G 结尾,表明首选密码子更倾向于以 U/A 结尾。其中 RSCU 值最高的密码子是编码精氨酸的 AGA,其次是编码亮氨酸的 UUA。见图 10。

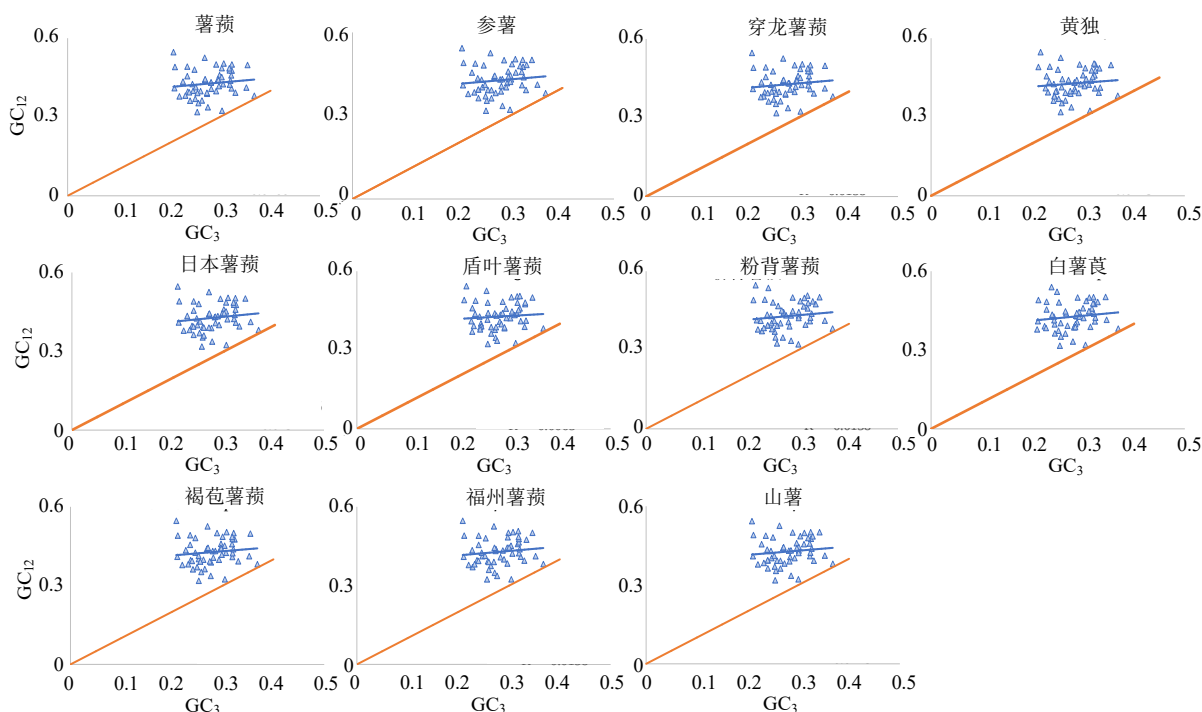


图 6 薯茛属植物叶绿体基因的密码子中性图谱分析

Fig. 6 Codon neutrality plot analysis of *Dioscorea* species chloroplast genes

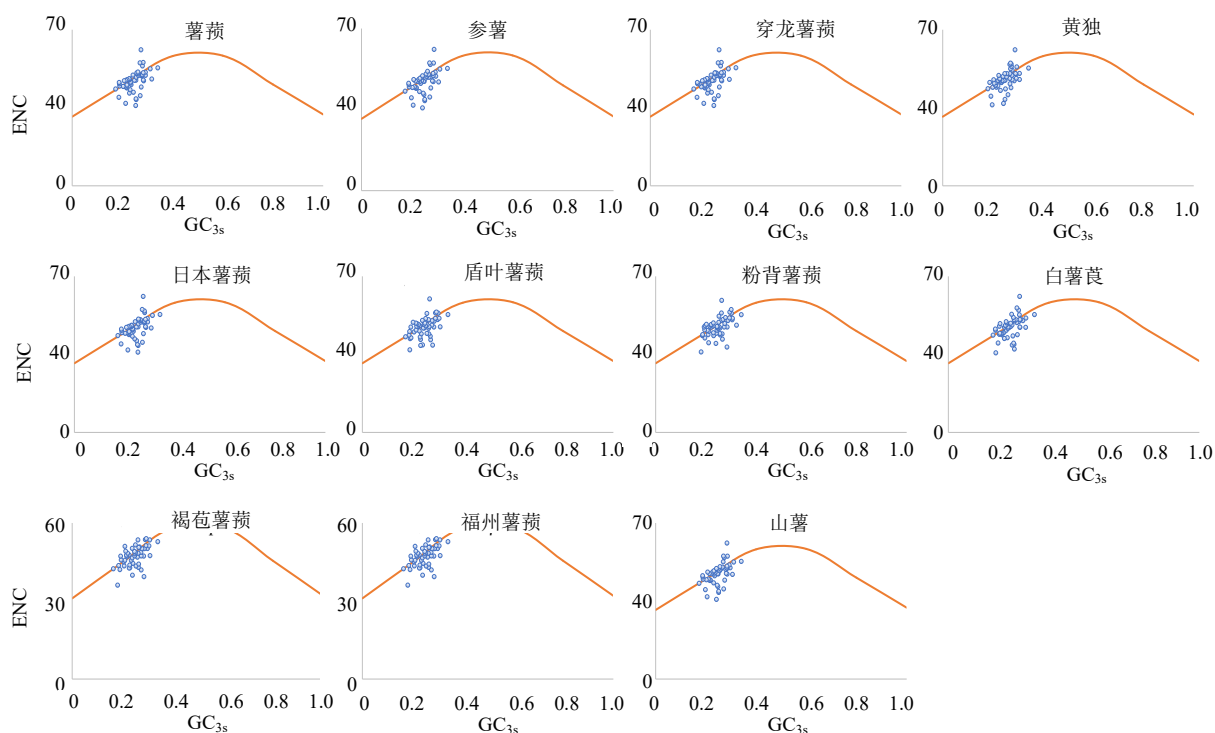


图 7 薯蓣属植物叶绿体基因组中 ENC 和 GC₃ 的关系 (ENC-plot 分析)

Fig. 7 Relationship between ENC and GC₃ in *Dioscorea* species chloroplast genomeS (ENC-plot analysis)

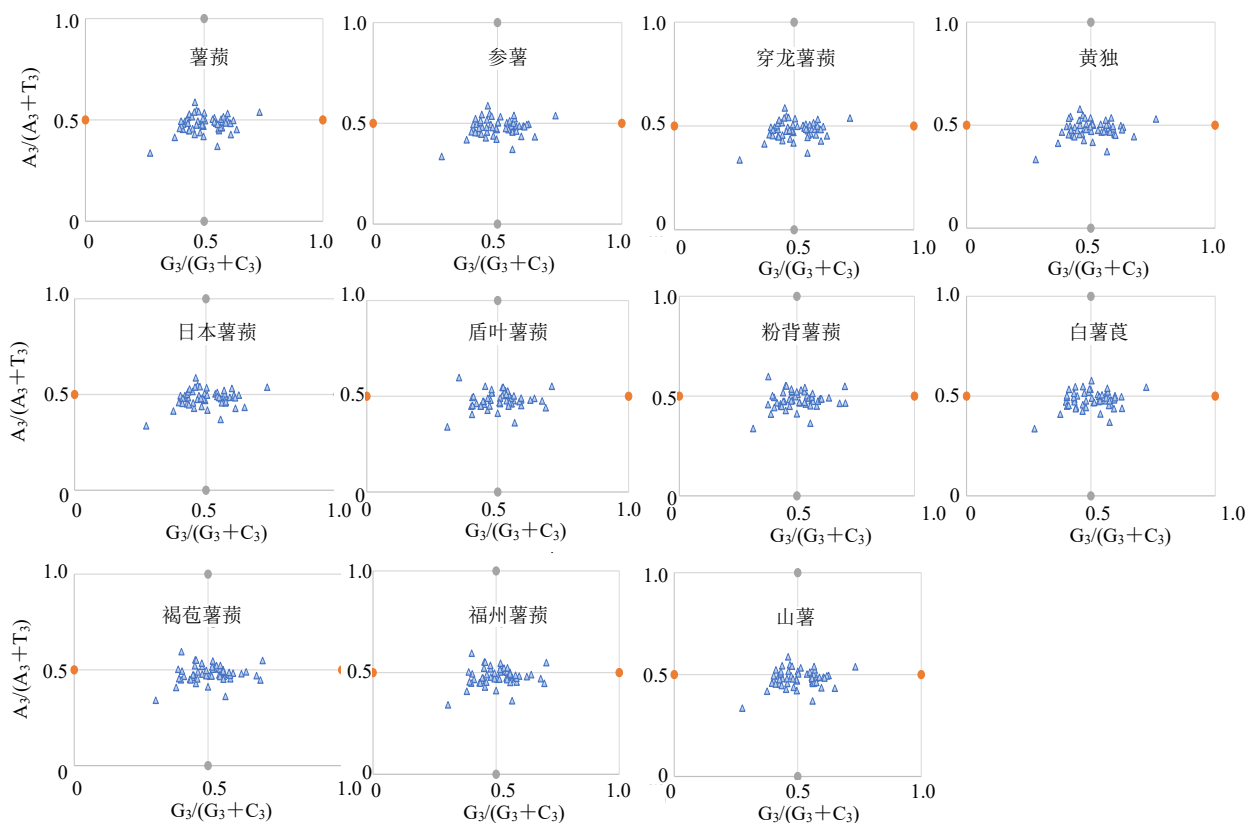


图 8 11 种薯蓣属植物的 PR2-plot 分析

Fig. 8 PR2 Plot analysis of 11 species of *Dioscorea* species

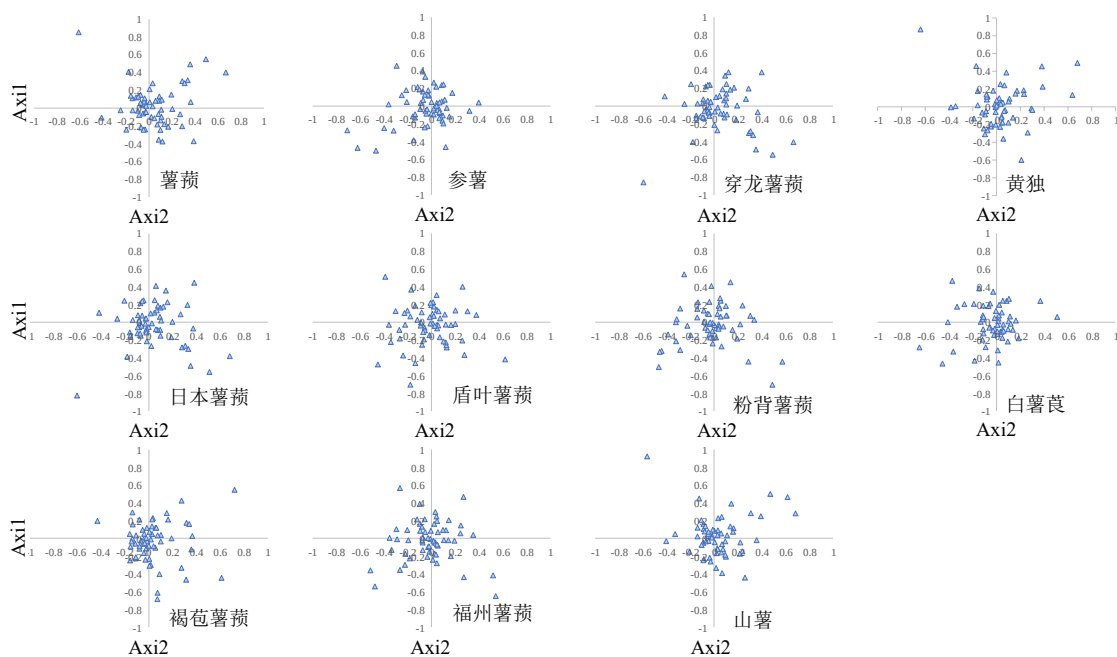


图 9 薯蓣属植物叶绿体基因组的 COA 分析

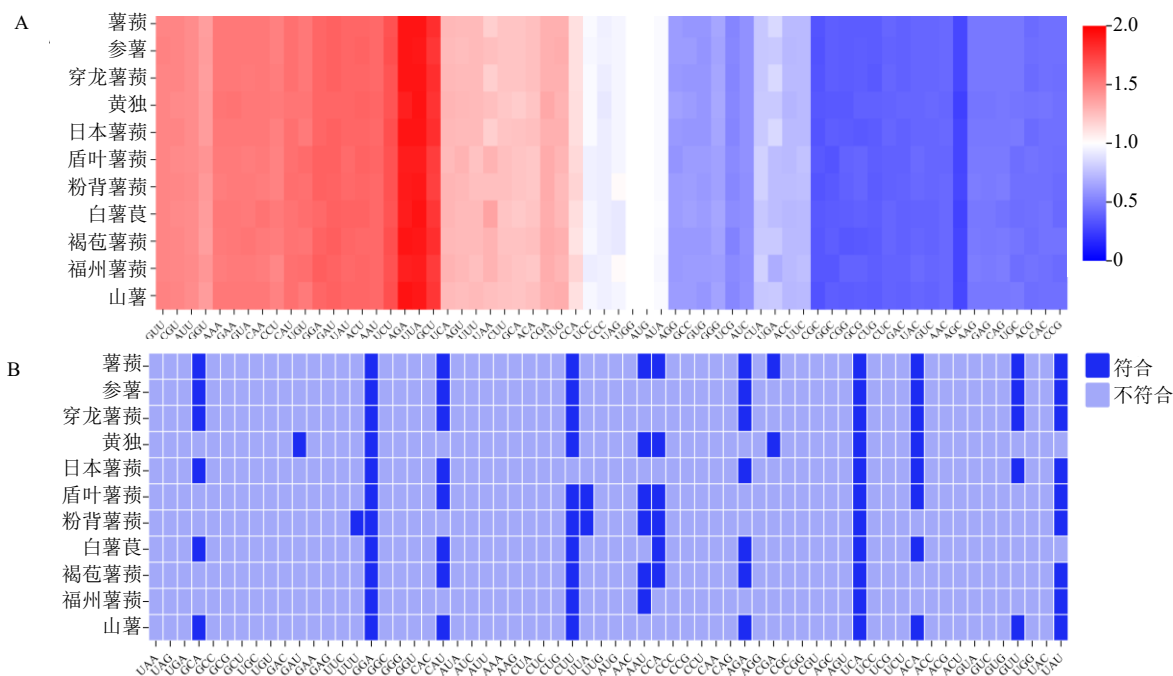
Fig. 9 COA analysis of chloroplast genomes in *Dioscorea* species深蓝色背景: $RSCU > 1$ 和 $\Delta RSCU \geq 0.08$; 浅蓝色背景: $RSCU > 1$ 。Dark blue background: $RSCU > 1$ and $\Delta RSCU \geq 0.08$; light blue background: $RSCU > 1$ 。

图 10 11 种植物 RSCU 热图 (A) 和最优密码子 (B)

Fig. 10 RSCU heatmap (A) and optimal codons (B) of 11 plant species

3.3 基于 DNA 条形码的系统发育分析

3.3.1 薯蓣属 *matK*、*rbcL*、*psbA-trnH* 序列特征及差异性分析 11 种植物共获 106 条序列。*rbcL* 序列

长 462~722 bp, GC 含量 44%~46%, 保守位点占比 100%, 变异位点为 0, 表明其长度和 GC 含量稳定, 保守性强; *matK* 序列长 707~870 bp, GC 含量

31%~32%，物种数、保守和变异位点最多，其中参薯、穿龙薯蓣、盾叶薯蓣分别有 1、2、3 个变异位点；*psbA-trnH* 序列长 246~503 bp，GC 含量 30%~37%，盾叶薯蓣和粉背薯蓣各有 1 个变异位点，其余植物为 0。3 种序列均具较高保守性。

3.3.2 3 种 DNA 条形码的鉴别效率分析 *psbA-trnH* 条形码序列分析显示，物种间最小遗传距离出现在日本薯蓣与参薯之间 (0.003 3)，最大距离为日本薯蓣与粉背薯蓣之间 (0.165 7)，而薯蓣与其他物种的遗传差异范围在 0.008 0~0.133 0。由于种间最小距离超过种内最大变异 (盾叶薯蓣，0.001 6) 见表 4，该序列可有效区分 8 种植物及薯蓣常见混淆种类。在 NJ 系统中，粉背薯蓣与穿龙薯蓣聚为一类，日本薯蓣自成一类，参薯则未形成独立分支，表明其与白薯蓣等物种遗传差异较小 (图 11-A)。

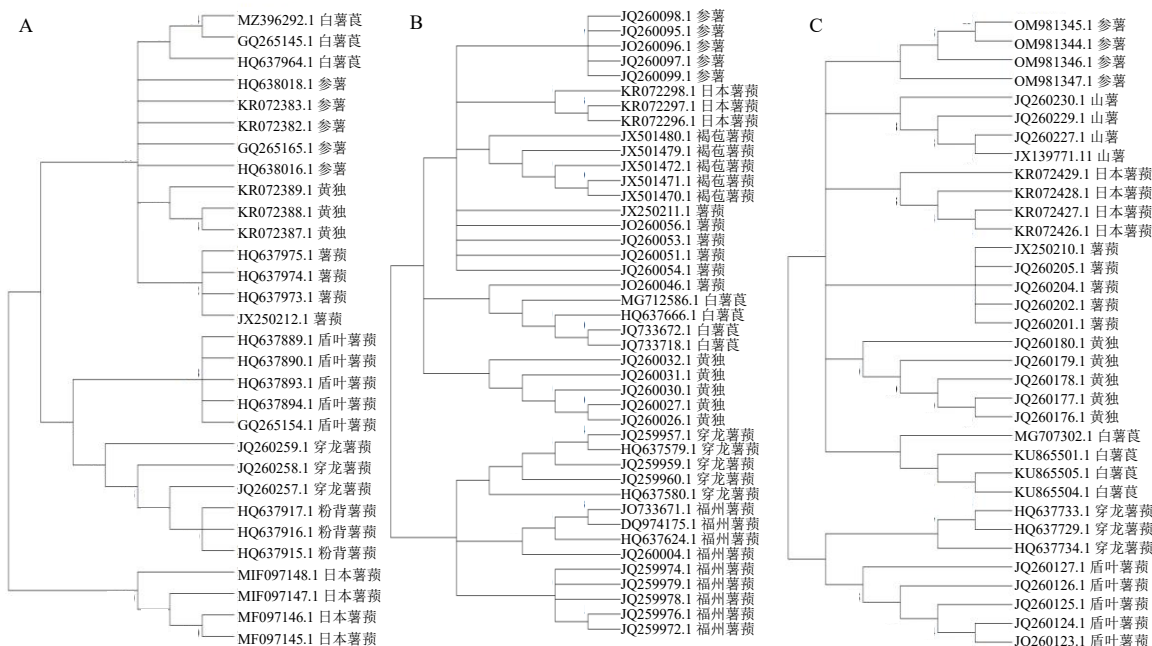
基于 *matK* 序列的分析表明，日本薯蓣与参薯的遗传距离最小 (0.003 2)，白薯蓣与穿龙薯蓣之间

最大 (0.065 0)，薯蓣与其他物种的遗传距离分布在 0.004 0~0.059 6 内。种间最小距离仍大于种内最大距离 (盾叶薯蓣，0.001 5) (表 4)，因此该标记可鉴别 9 类植物及薯蓣易混物种。NJ 树结构呈现两大分支：其中穿龙薯蓣、福州薯蓣和盾叶薯蓣聚集在一起；薯蓣未能单独分支，因其与参薯等种类差异较小；白薯蓣和黄独则形成高支持率 (99%~100%) 的独立分支 (图 11-B)。

在 *rbcl* 序列中，盾叶薯蓣与穿龙薯蓣的遗传距离最小 (0.004 0)，黄独与穿龙薯蓣之间最大 (0.027 7)，薯蓣与其他物种的遗传距离介于 0.004 2~0.020 2。由于所有种内遗传距离均为 0，种间最小距离仍大于种内变异，该序列成功区分出 8 种薯蓣属植物及其易混物种 (表 4)。NJ 进化树显示所有物种均形成独立分支，其中穿龙薯蓣与盾叶薯蓣聚为一簇，整体分为 2 大支系，种间鉴定支持率较高 (图 11-C)。

表 4 3 种 DNA 条形码遗传距离统计
Table 4 Genetic distance statistics of three DNA barcodes

条形码	种内K2P距离	种间K2P距离	薯蓣与其他物种距离
<i>rbcl</i>	0.000 0~0.000 0	0.004 0~0.027 7	0.004 2~0.020 2
<i>psbA-trnH</i>	0.000 0~0.001 6	0.003 3~0.165 7	0.008 0~0.133 0
<i>matK</i>	0.000 0~0.001 5	0.003 2~0.065 0	0.004 0~0.059 6



A-*psbA-trnH*; B-*matK*; C-*rbcl*.

图 11 基于条形码构建的邻接树

Fig. 11 NJ tree constructed based on barcode

3.3.3 基于机器学习的 DNA 条形码分析 如图 12 所示: BLOG 下, *rbcL*、*matK*、*psbA-trnH* 均达 100% 成功率。SMO 和 Naïve Bayes 分类器的鉴别成功率在 90%~100%, 其中 *rbcL* 的鉴别率最高, 可达 100%。Jrip 的鉴别成功率在 62%~100%,

psbA-trnH 的鉴别率最低。J48 的鉴别成功率在 34%~100%, 表现最差。总体而言, WEKA 分类器的鉴别成功率不及 BLOG, 但 *rbcL* 在各机器学习算法中均达 100% 成功率, 可作为薯蓣属鉴别的优选条形码。

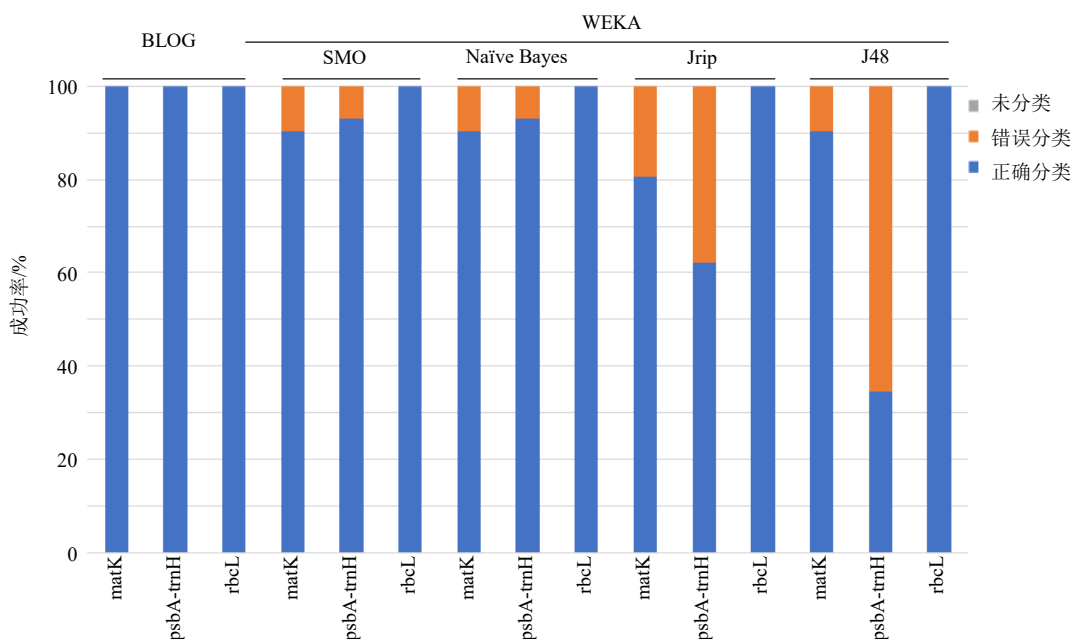


图 12 机器学习识别成功率柱状图

Fig. 12 Bar chart of machine learning recognition success rates

4 讨论

本研究发现 11 个薯蓣属物种叶绿体基因组在种间进化中高度保守, 与 Zhao 等^[21]对多种药用植物叶绿体基因组的比较结果一致, 均显示其结构稳定性较高; 然而, 薯蓣属物种在形态上呈现的多样性, 可能与线粒体或核基因组变异更为相关, 这一推断也与 Chen 等^[22]在禾本科植物中的多基因组演化规律相符。IR 边界的收缩扩张模式提示日本薯蓣与山薯、参薯亲缘较近, 该结果进一步通过最优密码子及系统发育树 (NJ 树) 验证, 与先前多项基于叶绿体基因片段 (如 *matK*+*rbcL*) 的系统发育重建结论基本一致^[23]。SSR 分析一致显示 A/T 碱基偏好, 且非编码区变异率显著高于编码区, 与已有研究在薯蓣非编码区序列中的高变异性报道吻合^[24], 提示这些区域可能在调控机制及适应性进化中发挥关键作用, 也为后续功能验证实验提供了方向。

密码子使用偏好性呈现显著的 A/T 偏向及第 3 位 A/U 结尾特征, 该模式与多种高等植物叶绿体基因组的报道一致^[19]。中性绘图、PR2-plot 及 ENC-plot 分析

一致指出自然选择是塑造密码子使用偏好的主要因素, 最优密码子为 GGA 和 UCA, 分别编码甘氨酸和丝氨酸。值得注意的是, 这 2 种氨基酸在一碳代谢及 UDP-葡萄糖基转移酶调控中起核心作用, 可能间接促进薯蓣皂苷的生物合成。该发现与之前皂苷合成通路中的关键酶基因密码子优化策略^[25]具有一致性, 说明跨物种保守机制可能普遍存在。今后可针对上述最优密码子开展合成生物学研究, 如构建优化版本的 UDP-葡萄糖基转移酶基因, 结合转基因及代谢组学手段, 验证其对皂苷积累的具体效应。

3 种常用 DNA 条形码 (*matK*、*psbA-trnH*、*rbcL*) 在本研究中均表现出良好的物种鉴别能力, 与之前多项植物分子鉴定研究结论一致^[26]。在系统发育拓扑结构上, *rbcL* 与 *matK* 显示穿龙薯蓣和盾叶薯蓣亲缘较近, 而 *psbA-trnH* 支持粉背薯蓣与穿龙薯蓣聚类, 叶绿体基因组系统发育则提示盾叶薯蓣、粉背薯蓣和福州薯蓣关系密切。该类拓扑差异在不同条形码中常见, 可能源于片段进化速率差异或谱系特异性进化事件, 与 Hollingsworth 等^[27]所强调的“多片段联合鉴

定”策略形成呼应。建议后续扩大样本量并结合核基因标记,构建更稳健的系统发育框架,尤其针对复合群和近缘种开展群体遗传学分析。

机器学习算法中,BLOG 表现优异,单一条形码鉴别成功率达 100%,而 WEKA 平台中 SMO 和 Naïve Bayes 分类器的鉴别率也超过 90%,与此前在其他植物类群中应用机器学习辅助物种鉴别的趋势一致(如 Carranza-Rojas 等^[28])。结果表明,机器学习方法可有效捕捉条形码序列中的鉴别特征,且对不同片段变异模式的适应性较强。建议后续开发融合多条条形码序列与形态特征数据的混合分类器,并引入更多解释性强的深度学习模型,以提升鉴别系统的鲁棒性和泛化能力。

综上,本研究从叶绿体基因组结构、密码子使用模式到分子鉴别技术,均与前人研究在不同植物类群中的发现存在较高一致性,说明这些进化与鉴别规律可能具有一定普适性;同时,结果也凸显出薯蓣属物种进化中的一些特异性事件,如特定条形码片段聚类冲突可能暗示不完全谱系分选或基因流事件。这些发现为后续开展功能验证实验、扩大样本类型及构建综合鉴别平台提供了扎实的理论依据和方法支撑。

利益冲突 所有作者均声明不存在利益冲突

参考文献

- [1] 朱绍杰,王宇飞,于丹,等. 分子生物学技术在薯蓣属植物中的应用研究进展 [J]. 中草药, 2024, 55(20): 7192-7202.
- [2] 郭璞注. 山海经 [M]. 上海: 上海古籍出版社, 1989.
- [3] 吴其潜. 植物名实图考 [M]. 上海: 商务印刷馆, 1957: 24.
- [4] 谢肇淛. 五杂俎 [M]. 北京: 京华出版社, 1989: 200.
- [5] 郑玉红,夏冰,杭悦宇,等. 山药原植物薯蓣及其近缘种的分子鉴别和亲缘关系研究 [J]. 南京农业大学学报, 2007, 30(2): 55-59.
- [6] 李伟,朱明玉,曾宇灵,等. 5 种黄精属植物叶绿体基因组从头组装及系统发育研究 [J]. 世界科学技术—中医药现代化, 2025, 27(1): 36-55.
- [7] 马双姣,周建国,李滢,等. 薯蓣和叉蕊薯蓣叶绿体基因组及特异性 DNA 条形码鉴定序列筛选研究 [J]. 中国科学: 生命科学, 2018, 48(5): 571-582.
- [8] 荣正普,王杰敏,裴林,等. 黄芩属药用植物叶绿体基因组密码子使用偏好性分析 [J]. 中草药, 2025, 56(1): 269-281.
- [9] Mahima K, Sunil Kumar K N, Rakesh K V, et al. Advancements and future prospective of DNA barcodes in the herbal drug industry [J]. *Front Pharmacol*, 2022, 13: 947512.
- [10] 何拓. 基于机器学习的黄檀属与紫檀属木材识别方法研究 [D]. 北京: 中国林业科学研究院, 2019.
- [11] 冯婷婷. 基于机器学习的三种药用植物 DNA 条形码分子鉴别研究 [D]. 广州: 广东药科大学, 2021.
- [12] 魏怡冰. 黔产薯蓣属药用植物资源调查、鉴别及亲缘关系研究 [D]. 贵阳: 贵阳中医学院, 2018.
- [13] Zheng M D, Zhang C, Ma R L, et al. Identification of Moslae Herba and its adulterants based on matK, ITS2 and its secondary structure [J]. *Chin J Mod Appl Pharm*, 2022, 39(17): 2222-2228.
- [14] 梁莹莹,吴文如,彭晓祺,等. 苹婆叶绿体全基因组序列特征及苹婆属系统发育分析 [J]. 中草药, 2024, 55(19): 6721-6729.
- [15] 李俊霖,郭淑红,张强,等. 蒙古黄芪叶绿体全基因组特征解析及亲缘性分析 [J]. 中草药, 2024, 55(7): 2366-2374.
- [16] 杜雨,李效雄,贾西贝,等. 枸杞叶绿体基因组密码子偏好性分析 [J]. 中草药, 2024, 55(4): 1316-1325.
- [17] Zhang J S, Feng M. Analysis of the Codon usage bias pattern in the chloroplast genomes of *Chloranthus* species (Chloranthaceae) [J]. *Genes*, 2025, 16(2): 186.
- [18] 徐荣,宋文政,乔振升,等. 富民枳叶绿体基因组密码子偏好性分析 [J]. 中草药, 2024, 55(20): 7093-7100.
- [19] Shen L W, Chen S Q, Liang M, et al. Comparative analysis of Codon usage bias in chloroplast genomes of ten medicinal species of Rutaceae [J]. *BMC Plant Biol*, 2024, 24(1): 424.
- [20] Yang Y, Liu X Y, He L X, et al. Comparative chloroplast genomics and Codon usage bias analysis in *Hevea* genus [J]. *Genes*, 2025, 16(2): 201.
- [21] Zhao Z Y, Wang X, Yu Y, et al. Complete chloroplast genome sequences of *Dioscorea*: Characterization, genomic resources, and phylogenetic analyses [J]. *PeerJ*, 2018, 6: e6032.
- [22] Chen J H, Hao Z D, Guang X M, et al. *Liriodendron* genome sheds light on angiosperm phylogeny and species-pair differentiation [J]. *Nat Plants*, 2019, 5(1): 18-25.
- [23] Xia W, Zhang B, Xing D, et al. Development of high-resolution DNA barcodes for *Dioscorea* species discrimination and phylogenetic analysis [J]. *Ecol Evol*, 2019, 9(18): 10843-10853.
- [24] Gong L, Ding X X, Guan W, et al. Comparative chloroplast genome analyses of *Amomum*: Insights into evolutionary history and species identification [J]. *BMC Plant Biol*, 2022, 22(1): 520.
- [25] Bureau J A, Oliva M E, Dong Y M, et al. Engineering yeast for the production of plant terpenoids using synthetic biology approaches [J]. *Nat Prod Rep*, 2023, 40(12): 1822-1848.
- [26] Pang X H, Song J Y, Zhu Y J, et al. Applying plant DNA barcodes for Rosaceae species identification [J]. *Cladistics*, 2011, 27(2): 165-170.
- [27] Hollingsworth P M, Graham S W, Little D P. Choosing and using a plant DNA barcode [J]. *PLoS One*, 2011, 6(5): e19254.
- [28] Carranza-Rojas J, Goeau H, Bonnet P, et al. Going deeper in the automated identification of Herbarium specimens [J]. *BMC Evol Biol*, 2017, 17(1): 181.

[责任编辑 时圣明]