

DOI:10.11913/PSJ.2095-0837.2022.20229

蔡元保, 杨祥燕. 澳洲坚果光壳种叶绿体基因组的密码子使用偏好性及其影响因素分析[J]. 植物科学学报, 2022, 40(2): 229-239

Cai YB, Yang XY. Codon usage bias and its influencing factors in the chloroplast genome of *Macadamia integrifolia* Maiden & Betche[J].*Plant Science Journal*, 2022, 40(2): 229-239

澳洲坚果光壳种叶绿体基因组的 密码子使用偏好性及其影响因素分析

蔡元保, 杨祥燕*

(广西壮族自治区农业科学院, 广西壮族自治区亚热带作物研究所, 南宁 530001)

摘要: 为确定澳洲坚果光壳种 (*Macadamia integrifolia* Maiden & Betche) 叶绿体基因组密码子偏好性形成的主要影响因素, 本研究通过其叶绿体基因组的 51 条蛋白编码序列, 系统分析其密码子的使用模式及其特征。密码子偏好性参数分析结果显示, 叶绿体基因密码子 3 位碱基的 GC 含量次序为 GC1 > GC2 > GC3; 有效密码子数 ENC 均值为 48.80 (> 阈值 45), 其密码子偏性较弱; GC3 与 ENC 呈显著相关, 影响其偏好性。中性绘图分析结果表明, GC12 与 GC3 的相关系数和回归曲线斜率分别为 0.186 和 0.265, 两者相关性不显著。ENC-plot 分析发现大部分基因接近于标准曲线, 其 ENC 比值位于 -0.05 ~ 0.05, 与期望值差异较小。PR2-plot 偏倚分析表明密码子第 3 位碱基的使用频率不均等, 嘧啶 T/C 高于嘌呤 A/G。RSCU 分析发现 29 个高频密码子、19 个高表达密码子及 16 个最优密码子中绝大多数偏好使用以 A 或 U 结尾。系统进化分析结果显示, 澳洲坚果光壳种和三叶澳洲坚果 (*M. ternifolia* F. Mueller) 的亲缘关系最近。研究结果说明, 作为主要因素的自然选择和碱基突变相对均衡地共同影响澳洲坚果光壳种叶绿体基因组较弱的密码子使用偏好性, 且影响其偏好性的密码子第 3 位碱基偏好使用 A 或 U。

关键词: 澳洲坚果; 叶绿体基因组; 密码子偏好性; 影响因素

中图分类号: Q943.2

文献标识码: A

文章编号: 2095-0837(2022)02-0229-11

Codon usage bias and its influencing factors in the chloroplast genome of *Macadamia integrifolia* Maiden & Betche

Cai Yuan-Bao, Yang Xiang-Yan*

(Guangxi Subtropical Crops Research Institute, Guangxi Academy of Agricultural Sciences, Nanning 530001, China)

Abstract: Codon usage bias of 51 genes from the chloroplast genome of *Macadamia integrifolia* Maiden & Betche was analyzed to determine the main factors affecting the formation of codon bias. The GC content at different codon positions in the chloroplast genes was GC1 > GC2 > GC3, and GC3 was significantly correlated with effective number of codons (ENC) (mean value 48.80 > threshold 45), indicating that the third codon position had direct impact on weak codon bias in the chloroplast genome of *M. integrifolia*. Neutral-plot analysis showed that there was no significant correlation between GC12 and GC3, with a correlation coefficient and regression slope of 0.186 and 0.265, respectively. ENC-plot analysis showed that most

收稿日期: 2021-09-23, 修回日期: 2021-11-24。

基金项目: 国家自然科学基金(31860537); 广西科技计划项目(桂科 AB19245008); 广西农业科学院基本科研业务专项(桂农科 2021YT154, 桂农科 2020YM56)。

This work was supported by grants from the National Natural Science Foundation of China (31860537), Guangxi Science and Technology Planning Project (GuikeAB19245008), and Basic Research Project of Guangxi Academy of Agricultural Sciences (Guinongke 2021YT154, Guinongke 2020YM56)。

作者简介: 蔡元保(1981-), 男, 硕士, 高级农艺师, 研究方向为植物生理与分子生物学(E-mail: caiyuanbao205@163.com)。

* 通讯作者(Author for correspondence. E-mail: yangxiangyan84412@126.com)。

genes were located around the standard curve, and their ENC ratios were distributed between -0.05 and 0.05 , suggesting that the differences between actual and expected ENC were small. PR2-plot analysis revealed that the third codon position was biased, where pyrimidine T/C was used more frequently than purine A/G. Relative synonymous codon usage (RSCU) analysis showed that most of the 29 high-frequency codons, 19 high-expression codons, and 16 optimal codons preferentially ended with A or U. Phylogenetic analysis showed that the relationship between *M. integrifolia* and *M. ternifolia* was similar. These results suggest that natural selection and base mutation are the main factors influencing weak codon bias in the chloroplast genome of *M. integrifolia*, and the third codon base, as another influencing factor, prefers to use A or U.

Key words: Macadamia (*Macadamia integrifolia*); Chloroplast genome; Codon usage bias; Influencing factors

相对于细胞核基因组, 叶绿体基因组较小、结构稳定, 具有基因拷贝数多、遗传信息相对保守等优势, 已被广泛用于植物适应性和多样性研究、系统演化分析、DNA 条码开发、遗传工程研究等^[1-3], 尤其是通过叶绿体基因工程实现外源基因的定点整合及高效稳定表达^[4]。

同义密码子在蛋白翻译中的不均衡使用造成叶绿体基因组存在密码子使用偏好性^[5, 6]。物种在长期进化中, 影响其密码子偏好性的主要因素是碱基突变和自然选择压力, 以及遗传漂移、基因序列长度及其表达水平、蛋白质疏水性、tRNA 丰度等其他因素^[7-9]。密码子偏好性可以通过调整翻译的效率和准确性从而影响基因的表达量; 且基因的密码子偏好性越强, 其表达量就越高^[9, 10]。此外, 不同物种的密码子使用模式还存在一定的差异^[11]。因此, 开展叶绿体基因组的密码子使用模式研究, 选用最优密码子改良外源基因和表达载体, 可以增强叶绿体遗传转化效率和外源基因的表达量, 对于植物重要性状的遗传改良及其叶绿体基因工程的推广应用具有重要的指导意义^[12, 13]。

随着多种植物叶绿体基因组被公开, 其密码子使用模式研究也日益增加, 如双子叶植物豌豆 (*Pisum sativum* L.)^[7]、刺榆 (*Hemiptelea davidii* (Hance) Planch.)^[14]、茄属 (*Solanum*)^[5] 植物等; 单子叶植物水稻 (*Oryza sativa* L.)^[8]、小麦 (*Triticum aestivum* L.)^[11], 及糜子 (*Panicum miliaceum* L.)^[15] 等黍属 (*Panicum*) 植物^[16]。澳洲坚果光壳种 (*Macadamia integrifolia* Maiden &

Betche) 原产于澳洲亚热带雨林, 属于山龙眼科澳洲坚果属, 且光壳种及其杂交种是澳洲坚果属最主要的栽培和食用种。虽然澳洲坚果属有多个物种的叶绿体基因组被公开, 但国内外对其基因组的深入研究并不多^[17-19]。目前, 基于澳洲坚果光壳种叶绿体基因组序列, 已明确了澳洲坚果属在真双子叶植物中的分子系统进化地位^[17, 20], 并揭示了澳洲坚果驯化的野生起源^[21]。但是, 国内外关于澳洲坚果属植物叶绿体基因组密码子使用模式研究尚未见报道。本研究通过分析澳洲坚果光壳种叶绿体基因组密码子的使用模式及其特征, 探讨其密码子使用偏好性的主要影响因素, 并筛选出最优密码子, 以期利用叶绿体基因工程改良澳洲坚果重要农艺性状提供科学依据。

1 材料与方法

1.1 研究材料

根据澳洲坚果叶绿体基因组 GenBank 登录号 (NC_025288.1), 从 NCBI (<https://www.ncbi.nlm.nih.gov/>) 数据库获取澳洲坚果光壳种叶绿体全基因组序列, 其全长为 159 714 bp, 其中共包含 88 条基因编码序列。为了减小重复序列和短序列变异所引起的分析误差, 采用 C 语言程序, 删除重复基因序列和长度 < 300 bp 的基因序列, 最终筛选出 51 个基因的编码区 CDS 序列应用在本研究的统计分析。

1.2 研究方法

1.2.1 密码子偏好性参数计算

通过 CodonW1.4.2 软件计算澳洲坚果光壳种 51 条基因序列的密码子偏好性参数, 包括有效密

密码子数 (Effective number of codons, ENC)、密码子适应指数 (Codon adaption index, CAI)、密码子偏好性指数 (Codon bias index, CBI)、最优密码子使用频率 (Frequency of optimal codons, FOP)、氨基酸长度 (Length of amino acid, Laa)、蛋白质疏水性 (Grand average of hydrophobicity, Gravy)、蛋白质芳香性 (Aromaticity of protein, Aromo)、同义密码子相对使用度 (Relative synonymous codon usage, RSCU) 等。通过 CUSP 软件计算 51 条基因序列的总 GC 含量及密码子第 1、2、3 位碱基 GC 含量, 分别用 GC_{all}、GC1、GC2、GC3 来表示。通过 SPSS 21.0 软件对这些参数进行相关性分析。

1.2.2 中性绘图分析

利用中性绘图分析来判断密码子第 1、2 与第 3 位碱基组成的相关性, 从而推测出密码子偏好性的影响因素。以各基因的 GC3 为横坐标, GC12 (GC1 和 GC2 的平均值) 为纵坐标, 绘制各基因的散点图及其回归曲线, 并分析 GC3 和 GC12 的相关性以确认密码子 3 个位置的碱基组成差异性。若相关性显著, 说明 3 个位置的碱基组成相似, 则碱基突变是其主要的影响因素; 若相关性不显著, 说明 3 个位置的碱基组成差异较大, 则偏好性受自然选择因素的影响更大^[7]。

1.2.3 ENC-plot 分析

通过 ENC-plot 分析各基因的密码子偏好情况及与碱基组成间的关系。采用 R 语言程序, 将各基因的 GC3 和 ENC 实际值分别作为横坐标和纵坐标, 构建两者的散点图, 并在图中根据标准曲线方程 $ENC = 2 + GC3 + 29/[GC3^2 + (1 - GC3)^2]$, 绘制标准曲线 (表示密码子偏好性仅由碱基组成决定)。通过散点位置即可判断该基因密码子偏好性的影响因素, 当散点在标准曲线上或附近时, 说明其偏好性受碱基突变因素的影响较大; 当散点远离标准曲线时, 说明其偏好性更多地受自然选择因素的影响^[8]。通过标准曲线计算 ENC 期望值, 并根据 ENC 比值 = (ENC 期望 - ENC 实际)/ENC 期望, 统计 ENC 比值频数及量化分析, 更好评估碱基突变和自然选择对密码子偏好性的影响程度。

1.2.4 PR2-plot 偏倚分析

通过 PR2-plot 偏倚分析, 探讨各基因密码子

第 3 位上的 A/T 和 G/C 之间突变的平衡情况。以 4 种同义密码子编码的氨基酸为分析对象, 通过 CodonW1.4.2 软件分析密码子第 3 位的碱基组成情况, 将 G3/(G3 + C3)、A3/(A3 + T3) 分别设为横坐标和纵坐标, 绘制平面图, 其中心点表示无偏倚时的密码子使用状态 (即 A = T 且 C = G), 各基因的散点与中心点的矢量距离表示其偏好性的方向和程度^[11]。

1.2.5 叶绿体基因组最优密码子筛选

将 51 个澳洲坚果光壳种叶绿体基因按照 ENC 值从小到大排序, 选择两端 10% 的基因 (各取 5 个基因), 分别构建高表达和低表达基因库, 并计算两基因库的 $\Delta RSCU$ 值。将 $\Delta RSCU \geq 0.08$ 的密码子定义为高表达密码子, 将 RSCU 值 > 1 的密码子定义为高频率密码子; 最后筛选出高表达密码子和高频率密码子的交集, 确定为澳洲坚果光壳种最优密码子^[5]。

1.2.6 基于叶绿体基因组的系统发育分析

从 GenBank 数据库中下载 12 种山龙眼目植物的叶绿体基因组序列, 包括莲属 (*Nelumbo*) 的黄连 (*N. lutea* Pers., 登录号 JQ336992.1) 和荷花 (*N. nucifera* Gaertn., 登录号 JQ336993.1)、悬铃木属 (*Platanus*) 的球悬铃木 (*P. occidentalis* L., 登录号 DQ923116.1)、泡花树属 (*Meliosma*) 的红柴枝 (*M. oldhamii* Maxim., 登录号 MZ491855.1)、清风藤属 (*Sabia*) 的小花清风藤 (*S. parviflora* Wall. ex Roxb., 登录号 MW566751.1) 和云南清风藤 (*S. yunnanensis* Franch., 登录号 NC_029431.1), 以及银桦属 (*Grevillea*) 的银桦 (*G. robusta* A. Cunn., 登录号 OK054586.1)、帝王花属 (*Protea*) 的乞力马扎罗帝王花 (*P. kilimandscharica* Engl., 登录号 MH362765.1)、山龙眼属 (*Helicia*) 的深绿山龙眼 (*H. nilagirica* Bedd., 登录号 NC_057271.1) 和瑞丽山龙眼 (*H. shweliensis* W. W. Smith, 登录号 NC_045942.1)、澳洲坚果属的澳洲坚果光壳种 (登录号 NC_025288.1) 和三叶澳洲坚果 (*M. ternifolia* F. Mueller, 登录号 NC_036416.1)。采用 ClustalX 2.1 软件对所有物种的叶绿体基因组序列进行比对, 采用 MEGA 7.0 软件的最大似然法 (Maximum likelihood, ML) 和邻接法 (Neighbor joining, NJ) 构建系统发育树。

2 结果与分析

2.1 起始密码子的使用模式分析

本研究发现,在澳洲坚果光壳种叶绿体基因组中,编码蛋白的基因共有 88 个,其中 83 个基因(占 94.32%)的起始密码子是 AUG;此外,基因 *psbL*、*rpl2* 和 *ndhD* 的起始密码子是 ACG,基因 *rpl16* 和 *rps19* 的起始密码子分别是 AUC 和 GUG。但蛋白翻译过程中起始密码子 ACG、AUC 和 GUG 无义,而是作为非起始密码子分别编码苏氨酸、异亮氨酸和缬氨酸。此外,其余遗传密码的使用情况与核基因一致。

2.2 密码子碱基组成和 ENC 值分析

密码子碱基组成和 ENC 值分析结果显示(表 1),51 个基因密码子的平均 GC 总含量为

38.95%,在 3 个位置上的 GC 含量次序为 GC1 (47.14%) > GC2 (39.97%) > GC3 (29.72%);除了基因 *matK*、*atpF*、*cemA* 和 *ycf2* 外,其余 47 个基因(占 92.16%)密码子的 GC1 和 GC2 明显高于 GC3。这说明密码子 3 个位置的碱基组成有差异(尤其是 GC 分布不均匀),第 3 位碱基富含 A 或 U。此外,澳洲坚果光壳种叶绿体基因的 ENC 值介于 39.78 ~ 59.39,均值为 48.80,有 43 个基因(占 84.31%)的 ENC 值高于阈值 45(此值作为区分密码子偏好性强弱的分界点)^[8],说明澳洲坚果光壳种叶绿体基因的密码子使用偏性普遍较弱。

2.3 密码子偏好性参数的相关性分析

密码子偏好性参数的相关性分析结果显示(表 2),GC_{all}与所有位置 GC 含量均呈极显著正相关,

表 1 澳洲坚果光壳种叶绿体基因组密码子的 GC 含量和 ENC 值
Table 1 GC content and ENC of chloroplast genome codons in *Macadamia integrifolia*

基因 Genes	GC1 /%	GC2 /%	GC3 /%	GC _{all} /%	ENC	基因 Genes	GC1 /%	GC2 /%	GC3 /%	GC _{all} /%	ENC
<i>psbA</i>	50.00	43.22	35.31	42.84	42.92	<i>rpl20</i>	37.29	44.07	22.88	34.75	43.06
<i>matK</i>	41.18	31.57	30.20	34.31	54.36	<i>clpP</i>	58.13	37.93	32.51	42.86	53.66
<i>atpA</i>	55.31	40.55	28.94	41.60	48.39	<i>psbB</i>	54.22	45.97	31.83	44.01	49.50
<i>atpF</i>	47.25	35.71	34.62	39.19	41.52	<i>petB</i>	48.61	41.67	31.48	40.59	41.84
<i>atpI</i>	51.21	37.50	27.02	38.58	47.21	<i>petD</i>	48.94	39.89	25.53	38.12	42.84
<i>rps2</i>	45.15	43.88	27.00	38.68	47.21	<i>rpoA</i>	47.11	34.35	29.48	36.98	50.79
<i>rpoC2</i>	46.24	38.64	31.18	38.69	51.37	<i>rps11</i>	54.96	54.96	26.72	45.55	48.31
<i>rpoC1</i>	50.95	37.92	27.23	38.70	49.23	<i>rps8</i>	42.11	42.86	30.83	38.60	49.99
<i>rpoB</i>	51.45	38.00	30.44	39.96	50.73	<i>rpl14</i>	54.47	38.21	27.64	40.11	54.49
<i>psbD</i>	52.82	43.50	33.33	43.22	46.20	<i>rpl16</i>	51.09	51.82	27.01	43.31	39.78
<i>psbC</i>	54.01	46.20	34.60	44.94	47.01	<i>rps3</i>	47.22	32.87	26.39	35.49	47.89
<i>psaB</i>	48.44	43.27	33.88	41.86	50.34	<i>rpl22</i>	39.31	41.38	25.52	35.40	44.87
<i>psaA</i>	52.73	43.68	33.69	43.36	48.90	<i>rpl2</i>	50.36	48.93	32.86	44.05	54.87
<i>ycf3</i>	47.93	39.05	30.77	39.25	54.54	<i>ycf2</i>	41.49	35.38	37.01	37.96	53.40
<i>rps4</i>	50.00	39.11	26.24	38.45	51.41	<i>ndhB</i>	42.07	39.53	30.92	37.51	46.41
<i>ndhJ</i>	50.31	40.25	34.59	41.72	59.39	<i>rps7</i>	53.21	45.51	23.08	40.60	45.85
<i>ndhK</i>	42.61	43.66	31.69	39.32	51.16	<i>ndhF</i>	36.76	35.81	25.00	32.52	45.77
<i>ndhC</i>	48.76	33.88	30.58	37.74	47.29	<i>ccsA</i>	34.16	38.20	29.19	33.85	47.03
<i>atpE</i>	50.75	42.54	31.34	41.54	48.85	<i>ndhD</i>	40.04	37.85	28.69	35.52	48.41
<i>atpB</i>	57.11	41.68	30.46	43.09	47.77	<i>ndhE</i>	38.61	34.65	28.71	33.99	54.38
<i>rbcL</i>	58.82	43.70	33.82	45.45	52.52	<i>ndhG</i>	45.20	35.03	25.99	35.40	48.52
<i>accD</i>	41.12	35.73	29.54	35.46	50.89	<i>ndhI</i>	41.44	38.12	29.28	36.28	48.79
<i>ycf4</i>	43.78	41.62	32.97	39.46	50.54	<i>ndhA</i>	44.51	40.66	25.82	37.00	45.26
<i>cemA</i>	38.70	30.43	32.17	33.77	48.28	<i>ndhH</i>	51.27	36.55	26.40	38.07	48.15
<i>petA</i>	52.65	36.45	31.15	40.08	55.37	<i>ycf1</i>	36.19	29.47	26.81	30.82	48.23
<i>rps18</i>	36.27	45.10	25.49	35.62	43.19	Mean	47.14	39.97	29.72	38.95	48.80

注: GC1、GC2、GC3 和 GC_{all} 分别表示密码子第 1、2、3 位碱基 GC 含量及总 GC 含量, ENC 表示有效密码子数。下同。

Notes: GC1, GC value of first codon position; GC2, GC value of second codon position; GC3, GC value of third codon position; GC_{all}, GC value of each gene; ENC, effective number of codons. Same below.

表 2 澳洲坚果光壳种叶绿体基因组主要参数的相关性分析
Table 2 Correlation analysis of main parameters in chloroplast genome of *Macadamia integrifolia*

参数 Parameters	GC1	GC2	GC3	GC _{all}	ENC	CAI	CBI	FOP	Laa	Gravy
GC2	0.390 **									
GC3	0.261	0.023								
GC _{all}	0.861 **	0.714 **	0.479 **							
ENC	0.162	-0.242	0.355 *	0.092						
CAI	0.419 **	0.045	0.428 **	0.404 **	-0.115					
CBI	0.411 **	0.196	0.331 *	0.440 **	-0.241	0.774 **				
FOP	0.394 **	0.212	0.384 **	0.455 **	-0.196	0.800 **	0.976 **			
Laa	-0.146	-0.288 *	0.264	-0.141	0.170	0.003	-0.099	-0.046		
Gravy	-0.016	-0.218	0.081	-0.088	-0.154	0.235	0.142	0.007	-0.144	
Aromo	-0.283 *	-0.295 *	0.291 *	-0.217	-0.053	0.269	0.096	0.078	0.169	0.539 **

注：CAI 为密码子适应指数，CBI 为密码子偏好性指数，FOP 为最优密码子使用频率，Laa 为氨基酸长度，Gravy 为蛋白质疏水性，Aromo 为蛋白质芳香性。*， $P < 0.05$ ；**， $P < 0.01$ 。
Notes: CAI, codon adaption index; CBI, codon bias index; FOP, frequency of optimal codons; Laa, length of amino acid; Gravy, grand average of hydropathicity; Aromo, aromaticity of protein.

GC1 与 GC2 呈极显著正相关，但 GC1、GC2 与 GC3 的相关性都不显著，表明澳洲坚果光壳种叶绿体密码子第 1、2 位的碱基组成相似，但与第 3 位存在明显差异。ENC 只与 GC3 呈显著正相关，表明密码子第 3 位碱基的 GC 含量直接影响其偏好性和基因的表达量。

相关性分析结果也表明(表 2)，CAI、CBI 及 FOP 三者之间呈极显著高度正相关；三者与 GC 含量指标(GC1、GC3 和 GC_{all})呈显著或极显著正相关，说明碱基组成(尤其是 GC 含量)影响叶绿体基因的表达水平；但三者与 ENC 的相关性都不显著，说明叶绿体基因的表达水平与密码子偏性无直接相关性。

此外，氨基酸长度(Laa)只与 GC2 呈显著负相关，表明密码子 3 个位置中第 2 位碱基组成影响氨基酸长度。氨基酸长度(Laa)与 ENC 相关性不显著，表明基因序列长度影响密码子的使用偏好性很小。Gravy 只与 Aromo 极显著相关，与其他参数都不相关；Aromo 还与 GC 含量指标(GC1、GC2 和 GC3)显著相关。

2.4 中性绘图分析

中性绘图分析表明(图 1)，各基因 GC12 值在 0.3283 ~ 0.5496，GC3 值在 0.2288 ~ 0.3701；仅 2 个基因(*cemA* 和 *ycf2*)坐落在对角线附近，其余基因(尤其是基因 *rps11*、*rpl16* 和 *rps7*)均坐落在对角线上方的较远位置。GC12 与 GC3 的相关系数只有 0.186，回归曲线斜率为 0.265(其突变

对偏好性的效应占 26.50%)，两者相关性不显著。说明大部分基因密码子第 3 位与第 1、2 位的碱基组成存在明显差异，其进化方式也有所不同。由此可见，除了突变压力的较弱作用外，澳洲坚果光壳种叶绿体基因组的密码子偏好性形成中主要受自然选择因素的影响，尤其是密码子第 3 位碱基。

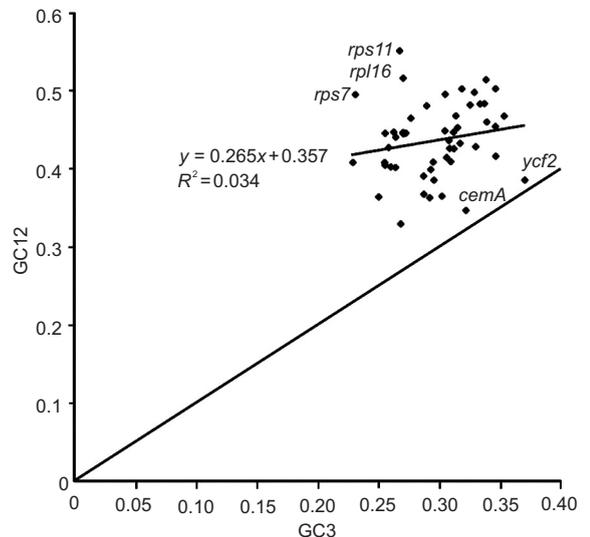


图 1 澳洲坚果光壳种叶绿体基因的中性绘图分析
Fig. 1 Neutral-plot analysis of chloroplast genes of *Macadamia integrifolia*

2.5 ENC-plot 分析

ENC-plot 分析发现大部分基因坐落在标准曲线上或其附近(图 2)，说明碱基突变对其密码子偏好性产生较大影响；少数基因(尤其是基因 *rpl16*、*petB*、*psbA* 和 *atpF*)远离标准曲线，说明其偏好

性的影响因素更多来自自然选择效应。在此基础上采用 ENC 比值频数分析(表 3), 可见 ENC 比值位于 -0.05 ~ 0.05 的基因有 31 个(占 60.8%), 在该范围内实际 ENC 值和期望 ENC 值的差异较小, 接近标准曲线。说明碱基突变是澳洲坚果光壳种叶绿体基因密码子偏性的重要影响因素, 但自然选择对其偏性也有较弱影响。

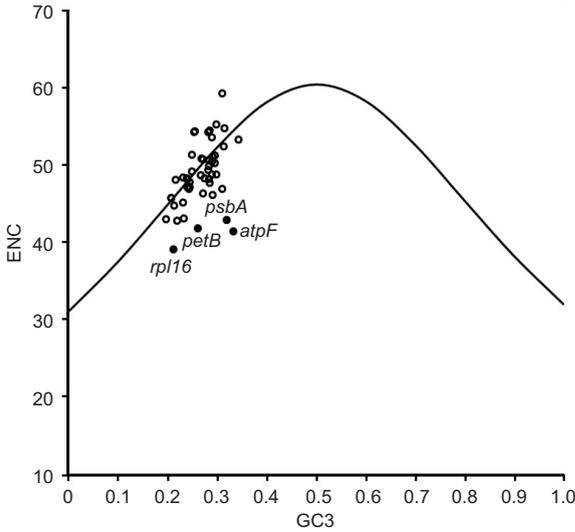


图 2 澳洲坚果光壳种叶绿体基因的 ENC-polt 绘图分析
Fig. 2 ENC-plot analysis of chloroplast genes of *Macadamia integrifolia*

表 3 ENC 比值频数分布
Table 3 Distribution of ENC ratio

组限 Class limits	中值 Mid-value	频数 Frequency No.	组频/% Frequency
-0.15 ~ -0.05	-0.10	7	13.7
-0.05 ~ 0.05	0	31	60.8
0.05 ~ 0.15	0.10	10	19.6
0.15 ~ 0.25	0.20	3	5.9
合计 Total		51	100.0

2.6 PR2-plot 偏倚分析

PR2-plot 偏倚分析结果显示(图 3), 各基因在 PR2 平面图 4 个区域的位点分布不均匀, 大多数基因坐落在图的下半部, 尤其是左下方区域, 表明该区域基因密码子第 3 位的碱基存在偏性(使用频率 T 高于 A, C 高于 G), 即密码子第 3 位嘧啶 T/C 的使用普遍比嘌呤 A/G 更频繁; 尤其图下半部的基因 *psbD*、*psbA*、*rps8* 和 *petB* 偏离中心点最

远。因此, 4 种碱基的使用频率不均等, 表明澳洲坚果光壳种叶绿体基因组的密码子偏好性受多种因素的综合作用, 受碱基突变影响的同时, 还受到自然选择等影响。

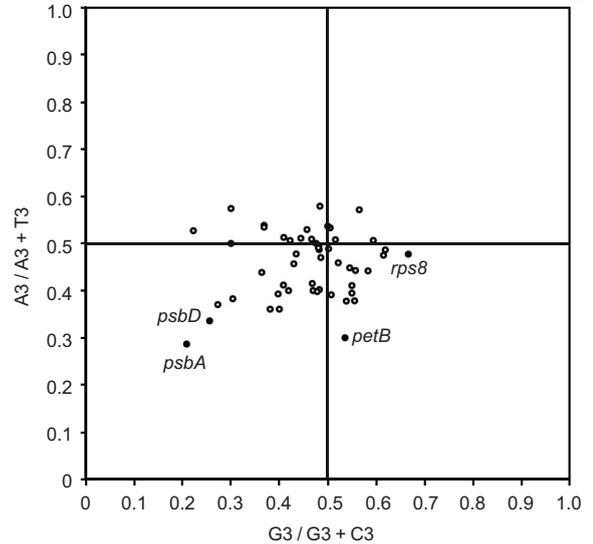


图 3 澳洲坚果光壳种叶绿体基因的 PR2-plot 分析
Fig. 3 PR2-plot analysis of chloroplast genes of *Macadamia integrifolia*

2.7 密码子 RSCU 分析及最优密码子筛选

各氨基酸的 RSCU 分析结果表明(表 4), 有 29 个密码子的 RSCU 值 > 1, 被确定为高频密码子, 其中除了 UUG 以 G 结尾外, 其余以 A 或 U 结尾的密码子占 96.55%(分别有 12 个和 16 个), 说明澳洲坚果光壳种叶绿体基因组偏好使用以 A 或 U 结尾的密码子。

根据 51 个基因密码子的 ENC 值分别构建高表达基因库(基因 *rpl16*、*atpF*、*petB*、*petD* 和 *psbA*)和低表达基因库(基因 *rpl14*、*ycf3*、*rpl2*、*petA* 和 *ndhJ*)。高、低表达基因库的 RSCU 分析发现有 19 个密码子的 $\Delta RSCU \geq 0.08$, 被确定为高表达密码子(有 84.21% 的密码子以 A 或 U 结尾)。同时满足高频密、高表达密码子的筛选标准, 最后评定 16 个澳洲坚果光壳种叶绿体基因组的最优密码子, 即 UUA、CUU、AUU、GUU、GUA、AGU、CCU、ACU、GCU、CAA、AAA、GAU、UGU、CGU、CGA、GGU。这些最优密码子全部以 A 或 U 结尾, 分别有 5 个和 11 个, 说明澳洲坚果光壳种叶绿体基因组的最优密码子第 3 位也偏好使用 A 或 U。

表 4 澳洲坚果光壳种叶绿体基因组 RSCU 分析及最优密码子确定

Table 4 RSCU analysis and putative optimal codons in chloroplast genome of *Macadamia integrifolia*

氨基酸 Amino acid	密码子 Codon	基因组 Genome		高表达基因 High expressed genes		低表达基因 Low expressed genes		Δ RSCU
		数目 No.	RSCU	数目 No.	RSCU	数目 No.	RSCU	
Phe	UUU	737	<u>1.18</u>	33	1.03	23	1.07	-0.04
	UUC	507	0.82	31	0.97	20	0.93	0.04
Leu	UUA ***	612	<u>1.60</u>	34	1.98	20	1.21	0.77
	UUG	514	<u>1.35</u>	27	1.57	25	1.52	0.05
	CUU *	439	<u>1.15</u>	18	1.05	15	0.91	0.14
	CUC	204	0.53	2	0.12	9	0.55	-0.43
	CUA	347	0.91	17	0.99	18	1.09	-0.10
	CUG	176	0.46	5	0.29	12	0.73	-0.44
Ile	AUU **	831	<u>1.34</u>	41	1.45	33	1.04	0.41
	AUC	463	0.75	21	0.74	27	0.85	-0.11
	AUA	566	0.91	23	0.81	35	1.11	-0.30
Val	GUU **	383	<u>1.29</u>	24	1.23	12	0.86	0.37
	GUC	175	0.59	5	0.26	9	0.64	-0.38
	GUA ***	434	<u>1.46</u>	41	2.10	21	1.50	0.60
	GUG	198	0.67	8	0.41	14	1.00	-0.59
Ser	UCU	395	<u>1.63</u>	23	2.00	18	1.96	0.04
	UCC *	236	0.97	12	1.04	8	0.87	0.17
	UCA	305	<u>1.26</u>	9	0.78	10	1.09	-0.31
	UCG	156	0.64	3	0.26	4	0.44	-0.18
	AGU ***	268	<u>1.11</u>	18	1.57	9	0.98	0.59
	AGC	95	0.39	4	0.35	6	0.65	-0.30
Pro	CCU ***	291	<u>1.36</u>	26	1.86	14	1.12	0.74
	CCC	195	0.91	11	0.79	9	0.72	0.07
	CCA	249	<u>1.16</u>	16	1.14	16	1.28	-0.14
	CCG	121	0.57	3	0.21	11	0.88	-0.67
Thr	ACU ***	337	<u>1.44</u>	30	1.97	8	0.82	1.15
	ACC	188	0.80	16	1.05	12	1.23	-0.18
	ACA	294	<u>1.25</u>	11	0.72	12	1.23	-0.51
	ACG	119	0.51	4	0.26	7	0.72	-0.46
Ala	GCU ***	412	<u>1.63</u>	40	1.93	19	1.25	0.68
	GCC	168	0.67	12	0.58	16	1.05	-0.47
	GCA	286	<u>1.13</u>	21	1.01	15	0.98	0.03
	GCG	142	0.56	10	0.48	11	0.72	-0.24
Tyr	UAU	615	<u>1.51</u>	21	1.14	26	1.21	-0.07
	UAC	199	0.49	16	0.86	17	0.79	0.07
His	CAU	370	<u>1.40</u>	10	1.05	22	1.47	-0.42
	CAC **	157	0.60	9	0.95	8	0.53	0.42
Gln	CAA ***	552	<u>1.40</u>	23	1.59	10	0.77	0.82
	CAG	235	0.60	6	0.41	16	1.23	-0.82
Asn	AAU	662	<u>1.47</u>	34	1.31	36	1.31	0
	AAC	241	0.53	18	0.69	19	0.69	0
Lys	AAA **	779	<u>1.42</u>	27	1.69	45	1.32	0.37
	AAG	320	0.58	5	0.31	23	0.68	-0.37

续表 4

氨基酸 Amino acid	密码子 Codon	基因组 Genome		高表达基因 High expression genes		低表达基因 Low expression genes		ΔRSCU
		数目 No.	RSCU	数目 No.	RSCU	数目 No.	RSCU	
Asp	GAU *	552	<u>1.52</u>	17	1.48	16	1.19	0.29
	GAC	172	0.48	6	0.52	11	0.81	-0.29
Glu	GAA	738	<u>1.44</u>	47	1.57	33	1.50	0.07
	GAG	284	0.56	13	0.43	11	0.50	-0.07
Cys	UGU ***	195	<u>1.35</u>	8	2.00	10	1.43	0.57
	UGC	94	0.65	0	0	4	0.57	-0.57
Trp	UGG	371	1.00	23	1.00	13	1.00	0
Arg	CGU ***	235	<u>1.10</u>	24	2.32	11	0.73	1.59
	CGC **	86	0.40	6	0.58	4	0.27	0.31
	CGA *	284	<u>1.33</u>	15	1.45	18	1.20	0.25
	CGG	134	0.63	3	0.29	14	0.93	-0.64
	AGA	379	<u>1.77</u>	12	1.16	28	1.87	-0.71
	AGG	167	0.78	2	0.19	15	1.00	-0.81
Gly	GGU ***	450	<u>1.31</u>	48	2.09	21	1.12	0.97
	GGC	163	0.47	10	0.43	9	0.48	-0.05
	GGA	501	<u>1.45</u>	24	1.04	30	1.60	-0.56
	GGG	264	0.77	10	0.43	15	0.80	-0.37

注: *, ΔRSCU ≥ 0.08; **, ΔRSCU ≥ 0.3; ***, ΔRSCU ≥ 0.5。高频密码子(即 RSCU > 1)用下划线标出; 最优密码子加粗表示。
Notes: High frequency codons are underlined (RSCU > 1); optimal codons are in bold.

2.8 基于叶绿体基因组的系统发育分析

基于叶绿体基因组的聚类分析结果如图 4 所示, 12 种山龙眼目植物都以 100% 的支持率被聚成两大分支, 清风藤科、莲科和悬铃木科聚成一支; 山龙眼科与这些科的遗传距离较远, 被单独聚

成一支, 其中与悬铃木科的亲缘关系最近。同属于山龙眼科的帝王花属、银桦属、山龙眼属和澳洲坚果属中, 澳洲坚果属和山龙眼属的遗传距离最近, 其亲缘关系最为紧密, 其中澳洲坚果光亮种和三叶澳洲坚果的亲缘关系最近。

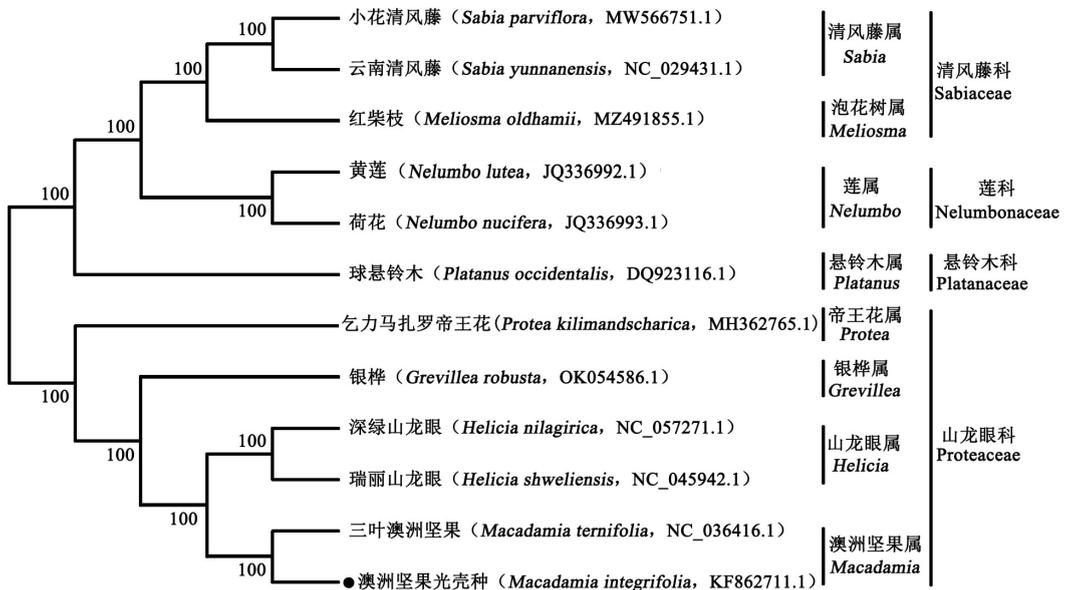


图 4 基于叶绿体基因组的 12 种山龙眼目植物系统发育树

Fig. 4 Phylogenetic tree of 12 Proteales species based on chloroplast genomes

3 讨论

密码子使用模式是基因组进化关系研究的主要指标之一。本研究中, ACG、AUC 和 GUG 是澳洲坚果光壳种叶绿体极少数基因的起始密码子, 在水稻、棉花 (*Gossypium herbaceum* L.)、小麦等植物中也有类似的起始密码子使用情况^[8, 11]。而这些密码子又是一些原核生物的起始密码子, 这又为叶绿体进化的内吞假说提供一个重要的佐证。在长期进化中, 同义密码子3个位置中只有第3位的碱基突变通常受环境的选择压力小, 其碱基组成(尤其是 GC 含量)影响着密码子偏好性^[6]。本研究表明澳洲坚果光壳种叶绿体基因中, 密码子3位碱基的 GC 含量依次为第1位 > 第2位 > 第3位; 尤其是第3位碱基的 GC 含量普遍低于前两位, 且富含 A 和 U, 直接影响其密码子偏好性。这和水稻^[8]、大花飞燕草 (*Delphinium grandiflorum* L.)^[22]、大戟科植物^[6]、黍属 (*Panicum*) 植物^[16] 等众多物种密码子碱基组成的变化趋势及使用模式极其相似, 进一步说明进化史中高等植物叶绿体基因组的密码子相对保守, 其使用规律高度相似。

影响不同物种密码子偏好性形成的多种因素中, 突变和选择是最为重要的两个驱动力^[7, 14, 23]。中性绘图和 ENC-plot 绘图综合分析表明, 碱基突变和自然选择这两大重要因素, 相对均衡且共同影响澳洲坚果光壳种叶绿体基因组较弱的密码子偏好性。而且, PR2-plot 分析的碱基使用不平衡结果也进一步证实了该结论, 这也符合“突变-选择-漂变”的理论观点^[24]。高等植物叶绿体基因组较弱的密码子偏好性中, 大多数植物受到突变和选择共同影响, 只是其作用大小不同而已; 只有少数植物与本研究的澳洲坚果光壳种一样, 受突变和选择的相对均衡影响, 如陆地棉 (*Gossypium hirsutum* L.)^[25]、凉粉草 (*Mesona chinensis* Benth.)^[26]、咖啡 (*Coffea arabica* L.)^[27]、菠萝 (*Ananas comosus* (L.) Merr.)^[28] 等。至于突变和选择相对均衡影响叶绿体基因组密码子偏性的分子作用机制还有待进一步研究。

高等种子植物核基因组的密码子使用模式研究表明, 大多数双子叶植物更趋向使用 A 或 U 结尾, 单子叶植物更偏好使用 G 或 C 结尾^[29]。本研究中澳洲坚果光壳种作为双子叶植物, 其叶绿体基因组

的密码子、高频密码子、高表达密码子和最优密码子都偏好使用 A 或 U 结尾, 和豌豆^[7]、刺榆^[14]、茄属植物^[5] 等多种双子叶植物相似, 说明这些双子叶植物叶绿体基因组的密码子使用模式与核基因组一样, 偏好使用 A 或 U 结尾。在蛋白翻译的过程中最优密码子能够有效提高其速率与精准度, 由此诱导基因的高效表达^[3, 9, 30]。本研究筛选出澳洲坚果光壳种叶绿体基因组的最优密码子有 16 个, 与大多数藻类和陆生植物的最优密码子偏好 NNA 型、NNU 型的使用模式相符^[31]; 且最优密码子数量较多, 其偏好性主要受强正向选择和突变压力的共同作用^[16, 32]。

与其他物种相比, 澳洲坚果光壳种和三叶澳洲坚果的亲缘关系最近, 在叶绿体基因组进化上具有类似的进化位置与历程, 推测它们具有相似的密码子使用偏好性。光壳种是澳洲坚果属中最主要的栽培和食用种, 其他近缘种可作为光壳种的杂交育种对象。因此, 根据本研究的澳洲坚果光壳种叶绿体基因组密码子的使用偏好性及其最优密码子, 可为澳洲坚果遗传育种及叶绿体基因工程的外源基因高效、稳定表达提供科学依据。

参考文献:

- [1] Xie DF, Yu Y, Deng YQ, Li J, Liu HY, et al. Comparative analysis of the chloroplast genomes of the Chinese endemic genus *Urophyssa* and their contribution to chloroplast phylogeny and adaptive evolution[J]. *Int J Mol Sci*, 2018, 19(7): 1847.
- [2] Meucci S, Schulte L, Zimmermann HH, Stoof-Leichsenring KR, Epp L, et al. Holocene chloroplast genetic variation of shrubs (*Alnus alnobetula*, *Betula nana*, *Salix* sp.) at the siberian tundra-taiga ecotone inferred from modern chloroplast genome assembly and sedimentary ancient DNA analyses[J]. *Ecology Evol*, 2021, 11(5): 2173–2193.
- [3] Wang K, Cui Y, Wang Y, Gao Z, Liu T, et al. Chloroplast genetic engineering of a unicellular green alga *Haemato-coccus pluvialis* with expression of an antimicrobial peptide[J]. *Mar Biotechnol*, 2020, 22(4): 572–580.
- [4] Wannathong T, Waterhouse JC, Young RE, Economou CK, Purton S. New tools for chloroplast genetic engineering allow the synthesis of human growth hormone in the green alga *Chlamydomonas reinhardtii*[J]. *Appl Microbiol Biotechnol*, 2016, 100(12): 5467–5477.
- [5] Zhang R, Zhang L, Wang W, Zhang Z, Du H, et al. <http://www.plantscience.cn>

- Differences in codon usage bias between photosynthesis-related genes and genetic system-related genes of chloroplast genomes in cultivated and wild *Solanum* species [J]. *Int J Mol Sci*, 2018, 19(10): 3142.
- [6] Wang Z, Xu B, Li B, Zhou Q, Wang G, et al. Comparative analysis of codon usage patterns in chloroplast genomes of six Euphorbiaceae species [J]. *PeerJ*, 2020, 8: e8251.
- [7] Bhattacharyya D, Uddin A, Das S, Chakraborty S. Mutation pressure and natural selection on codon usage in chloroplast genes of two species in *Pisum* L. (Fabaceae: Faboideae) [J]. *Mitochondrial DNA A DNA Mapp Seq Anal*, 2019, 30(4): 664–673.
- [8] Chakraborty S, Yengkhom S, Uddin A. Analysis of codon usage bias of chloroplast genes in *Oryza* species: codon usage of chloroplast genes in *Oryza* species [J]. *Planta*, 2020, 252(4): 67.
- [9] Zhao F, Zhou Z, Dang Y, Na H, Adam C, et al. Genome-wide role of codon usage on transcription and identification of potential regulators [J]. *Proc Natl Acad Sci USA*, 2021, 118(6): e2022590118.
- [10] Zhou Z, Dang Y, Zhou M, Li L, Yu CH, et al. Codon usage is an important determinant of gene expression levels largely through its effects on transcription [J]. *Proc Natl Acad Sci USA*, 2016, 113(41): e6117–e6125.
- [11] Tian G, Li G, Liu Y, Liu Q, Wang Y, et al. Polyploidization is accompanied by synonymous codon usage bias in the chloroplast genomes of both cotton and wheat [J]. *PLoS One*, 2020, 15(11): e0242624.
- [12] Wu S, Xu L, Huang R, Wang Q. Improved biohydrogen production with an expression of codon-optimized *hemH* and *lba* genes in the chloroplast of *Chlamydomonas reinhardtii* [J]. *Bioresour Technol*, 2011, 102(3): 2610–2616.
- [13] Kwon KC, Chan HT, León IR, Williams-Carrier R, Barkan A, et al. Codon optimization to enhance expression yields insights into chloroplast translation [J]. *Plant Physiol*, 2016, 172(1): 62–77.
- [14] Liu H, Lu Y, Lan B, Xu JJ. Codon usage by chloroplast gene is bias in *Hemiptelea davidii* [J]. *J Genet*, 2020, 99: 8.
- [15] 刘慧, 王梦醒, 岳文杰, 邢光伟, 葛玲巧, 等. 糜子叶绿体基因组密码子使用偏性的分析 [J]. 植物科学学报, 2017, 35(3): 362–371.
- Liu H, Wang MX, Yue WJ, Xing GW, Ge LQ, et al. Analysis of codon usage in the chloroplast genome of Broomcorn millet (*Panicum miliaceum* L.) [J]. *Plant Science Journal*, 2017, 35(3): 362–371.
- [16] Li G, Zhang L, Xue P. Codon usage pattern and genetic diversity in chloroplast genomes of *Panicum* species [J]. *Gene*, 2021, 802: 145866.
- [17] Nock CJ, Baten A, Barkla BJ, Furtado A, Henry RJ, et al. Genome and transcriptome sequencing characterises the gene space of *Macadamia integrifolia* (Proteaceae) [J]. *BMC Genomics*, 2016, 17(1): 937.
- [18] Liu J, Niu YF, Ni SB, He XY, Shi C. Complete chloroplast genome of a subtropical fruit tree *Macadamia ternifolia* (Proteaceae) [J]. *Mitochondrial DNA B Resour*, 2017, 2(2): 738–739.
- [19] Liu J, Niu YF, Ni SB, He XY, Zheng C, et al. The whole chloroplast genome sequence of *Macadamia tetraphylla* (Proteaceae) [J]. *Mitochondrial DNA B Resour*, 2018, 3(2): 1276–1277.
- [20] Nock CJ, Baten A, King GJ. Complete chloroplast genome of *Macadamia integrifolia* confirms the position of the Gondwanan early-diverging eudicot family Proteaceae [J]. *BMC Genomics*, 2014, 15(S9): S13.
- [21] Nock CJ, Hardner CM, Montenegro JD, Ahmad Termizi AA, Hayashi S, et al. Wild origins of macadamia domestication identified through intraspecific chloroplast genome sequencing [J]. *Front Plant Sci*, 2019, 10: 334.
- [22] Duan H, Zhang Q, Wang C, Li F, Tian F, 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.
- [23] Li G, Pan Z, Gao S, He Y, Xia Q, et al. Analysis of synonymous codon usage of chloroplast genome in *Porphyra umbilicalis* [J]. *Genes Genomics*, 2019, 41(10): 1173–1181.
- [24] Bulmer M. The selection-mutation-drift theory of synonymous codon usage [J]. *Genetics*, 1991, 129(3): 897–907.
- [25] 尚明照, 刘方, 华金平, 王坤波. 陆地棉叶绿体基因组密码子使用偏性的分析 [J]. 中国农业科学, 2011, 44(2): 245–253.
- Shang MZ, Liu F, Hua JP, Wang KB. Analysis on codon usage of chloroplast genome of *Gossypium hirsutum* [J]. *Scientia Agricultura Sinica*, 2011, 44(2): 245–253.
- [26] Tang D, Wei F, Cai Z, Wei Y, Khan A, et al. Analysis of codon usage bias and evolution in the chloroplast genome of *Mesona chinensis* Benth [J]. *Dev Genes Evol*, 2021, 231(1–2): 1–9.
- [27] Nair RR, Nandhini MB, Monalisha E, Murugan K, Sethuraman T, et al. Synonymous codon usage in chloroplast genome of *Coffea arabica* [J]. *Bioinformation*, 2012,

- 8(22): 1096–1104.
- [28] 杨祥燕, 蔡元保, 谭秦亮, 覃旭, 黄显雅, 等. 菠萝叶绿体基因组密码子偏好性分析[J]. *热带作物学报*, 2022, 43(3): 439–446.
Yang XY, Cai YB, Tan QL, Qin Xu, Huang XY, *et al.* Analysis of codon usage bias in the chloroplast genome of *Ananas comosus*[J]. *Chinese Journal of Tropical Crops*, 2022, 43(3): 439–446.
- [29] Wang L, Roossinck MJ. Comparative analysis of expressed sequences reveals a conserved pattern of optimal codon usage in plants[J]. *Plant Mol Biol*, 2006, 61(4–5): 699–710.
- [30] Quax TE, Claassens NJ, Söll D, Oost J. Codon bias as a means to fine-tune gene expression[J]. *Mol Cell*, 2015, 59(2): 149–161.
- [31] Qi YY, Xu WJ, Xing T, Zhao MM, Li NN, *et al.* Synonymous codon usage bias in the plastid genome is unrelated to gene structure and shows evolutionary heterogeneity[J]. *Evol Bioinform Online*, 2015, 11(1): 65–77.
- [32] Paul P, Malakar AK, Chakraborty S. Codon usage and amino acid usage influence genes expression level[J]. *Genetica*, 2018, 146(1): 53–63.

(责任编辑: 周媛)