两种茶树全基因组数据的密码子偏好性比较分析

王占军#* 李豹# 姜行舟 欧祖兰 徐忠东 代欢欢

(合肥师范学院生命科学学院, 合肥 230601)

摘要 研究密码子偏好性有利于提高两种茶树外源基因表达水平及探讨遗传进化规律。该研究综合运用CodonW 1.4.2、SPSS 22.0软件以及Perl、Python语言, 系统分析两种茶树基因组CDS序列, 从而得出密码子偏好模式和高频密码子, 明确密码子使用的变异来源。结果表明: 两种茶树基因组的密码子偏好性主要受突变压力影响, 在一定程度上也受到自然选择等因素的影响; 两种茶树基因组倾向于使用以A/T结尾的密码子, 且它们在密码子使用模式上具有较高的相似性; “云抗10号”(*Camellia sinensis* var. *assamica*, CSA)茶树基因的异源表达受体优先选择拟南芥、毛果杨和烟草, “舒茶早”(*Camellia sinensis* var. *assmica*, CSS)茶树基因的异源表达受体首选毛果杨。上述结果为分析两种茶树的进化规律、提高它们基因组中的基因异源表达效率提供了重要依据。

关键词 茶树; 全基因组; 密码子偏好性; 异源表达

Comparative Analysis of the Codon Preference Patterns in Two Species of *Camellia sinensis* Based on Genome Data

Wang Zhanjun#*, Li Bao#, Jiang Xingzhou, Ou Zulan, Xu Zhongdong, Dai Huanhuan

(*College of Life Science, Hefei Normal University, Hefei 230601, China*)

Abstract Analysis of codon preference patterns is helpful to improve the expression level of exogenous genes and explore the genetic evolution laws. In this paper, CodonW 1.4.2, SPSS 22.0 software and Perl, Python languages were used to systematically analyze the CDS sequences of genomes of the two species of *Camellia sinensis*, thus the pattern of codon preference and the high frequency codons were obtained, the source of codon usage variation were determined. The results showed that the codon preference of genomes of the two species of *Camellia sinensis* was mainly affected by mutation pressure, natural selection and other factors to some extent. It showed that the genomes of two species of *Camellia sinensis* preferred to use codons ending in A/T, and they had high similarity in codon usage patterns. It showed that heterologous expression receptors of *Camellia sinensis* var. *assamica* (CSA) genes were preferentially selected *Arabidopsis thaliana*, *Poplar* and *Tobacco*, but *Camellia sinensis* var. *assmica* (CSS) genes were preferentially selected *Poplar*. These results provide an important basis for analyzing the evolution of two species of *Camellia sinensis* and improving the efficiency of gene heterologous expression.

Keywords *Camellia sinensis*; whole genome; codon usage bias; heterologous expression

收稿日期: 2018-08-27 接受日期: 2018-10-12

安徽省自然科学基金面上项目(准号: 1708085MC76)、安徽省高校自然科学基金重点项目(批准号: KJ2015A186)和合肥师范学院教学研究项目(批准号: 2017jy43)资助的课题

#共同第一作者

*通讯作者。Tel: 0551-63674150, E-mail: wangzhanjunhxj@163.com

Received: August 27, 2018 Accepted: October 12, 2018

This work was supported by the General Project of Natural Science Foundation of Anhui Province (Grant No.1708085MC76), Key Project of Natural Science Foundation of Universities in Anhui Province (Grant No.KJ2015A186) and Teaching Research Project of Hefei Normal University (Grant No.2017jy43)

#These authors contributed equally to this work

*Corresponding author. Tel: +86-551-63674150, E-mail: wangzhanjunhxj@163.com

网络出版时间: 2018-11-27 16:57:12 URL: http://kns.cnki.net/kcms/detail/31.2035.Q.20181127.1649.002.html

密码子是mRNA分子中的一系列三联体核苷酸, 目前有64个通用密码子, 这些密码子共编码20种不同的氨基酸, 故密码子具有简并性[1]。例如, 在20种氨基酸中仅甲硫氨酸(Met)和色氨酸(Trp)分别对应1个密码子, 其他所有氨基酸均由2~6个不同的同义密码子编码。在翻译过程中, 同义密码子的使用频率不同, 即某一物种在编码基因时通常偏好于使用一种或几种特定的同义密码子, 这一密码子使用特性被称为密码子偏好性[2]。不同生物体的密码子偏好性差异很大, 甚至同一个生物体在编码不同基因时密码子的使用偏好也存在很大的差异, 分析可能由于基因组以及基因进化过程中承受的压力不同所致[3]。通过分析密码子使用偏好性, 不仅可以揭示相关物种间或某一物种的基因家族间的进化规律, 还可以优化外源基因的密码子, 这对于提高外源基因在宿主中的表达效率具有重要意义[4]。近年来, 通过密码子优化来提高外源基因表达效率的研究报道较多[5-7]。

随着高通量测序技术的快速发展, 大量物种的全基因组数据相继公布, 这也为人类开展全基因组密码子偏好性研究提供了强有力的保障。截至目前, 已有多篇在高等植物全基因组层面上证明物种内及物种间广泛存在同义密码子偏好性的报道, 如挪威云杉(*Picea abies*)和白云杉(*Picea glauca*)[8], 苹果(*Malus*×*domestica*)[9]及桑树(*Morus notabilis*)[10]等。两种云杉的基因组密码子分析[8]揭示, 它们的基因表达模式与其密码子的偏好性有关, 自然选择在两种云杉蛋白编码基因的翻译效率上起着重要的调控作用; 同时, 两种云杉的密码子偏好性与其基因家族的大小相关, 例如, 与单拷贝基因家族相比, 大量高拷贝基因的密码子偏好性较低。Li等[9]在针对苹果全基因组密码子偏好性分析时发现, 苹果密码子使用模式受到突变压力和自然选择的共同影响, 编码苹果蛋白的基因更倾向于使用以A/T结尾的密码子。Wen等[10]通过分析桑树与其他12种蔷薇科植物基因组的密码子偏好性, 揭示桑树的密码子使用模式受到编码基因的长度、碱基组成、突变压力、自然选择、基因表达水平和基因功能的共同影响; 其中, “基因的突变压力”和“翻译水平上的自然选择”是影响桑树基因组密码子偏好性的主要因素。

茶树[*Camellia sinensis* (L.) O. Kuntze]是被子植物山茶科(Theaceae)的代表成员之一, 是原产于中国西南部的特有种, 现已广泛种植于世界各地[11]; 因其叶中富含大量对人类有保健功效和药用价值的次生代谢物成分(如茶多酚、咖啡因、茶氨酸和维生素等)而备受青睐[12]。目前, 关于茶树密码子的偏好性研究仅局限于单基因的分析, 如茶树*CBF1*[3]、*ICE1*[13]、*SAD*[14]、*GPAT*[15]和*SPDS*[16]基因。2017年5月, 高立志研究组[11]首次公布了栽培茶树大叶茶种“云抗10号”(*Camellia sinensis* var. *assamica*, CSA)的全基因组序列。时隔一年, 宛晓春团队[12]公布了国家级茶树品种“舒茶早”(*Camellia sinensis* var. *assmica*, CSS)的全基因组草图。 因此, 本研究在基于目前公布的两种茶树全基因组数据的基础上开展密码子偏好性研究, 拟使用CodonW 1.4.2、SPSS 22.0软件及Perl、Python语言, 系统分析两种茶树全基因组的CDS序列, 旨在解析密码子偏好模式和高频密码子, 揭示密码子使用的变异来源。上述分析将有利于人们理解两种茶树的密码子使用模式, 同时为两种茶树提供合适的基因异源表达受体系统, 为通过优化密码子来提高茶树基因的表达量等研究提供重要的理论基础。

1 材料与方法

1.1 茶树CDS序列来源

从两种茶树全基因组数据网址(www.plantkingdomgdb.com/tea_tree/、pcsb.ahau.edu.cn:8080/CSS/)[11-12]中, 分别下载CSA的CDS序列[11](CSA的CDS数: 36 951)和CSS的CDS序列[12](CSS的CDS数: 33 932)。为了避免采样偏差, 从全基因组数据中筛选待分析的CDS序列时遵循如下原则[17]: (1)编码序列完整的蛋白质基因; (2)利用Perl语言删去长度<300 bp的CDS序列; (3)只在细胞质中翻译的基因。最终, 在CSA和CSS全基因组CDS序列中, 共筛选出符合3项筛选原则的CDS数分别为31 154条和33 917条。

1.2 密码子使用指数

应用Python语言计算两种茶树基因组CDS序列的密码子GC1、GC2、GC3和3个位置平均GC百分率。使用CodonW 1.4.2程序分析两种茶树基因组CDS序列中的密码子使用模式[18], 具体分析项目包括: 有效密码子数(effective number of codons, ENc)、同义密码子相对使用度(relative synonymous codon usage, RSCU)、相对同义密码子使用频率(relative frequency of synonymous codon, RFSC)、T3s、C3s、A3s、G3s和密码子适应指数(codon adaptation index,

CAI)、密码子偏性指数(codon bias index, CBI)、最优密码子使用频率(frequency of optimal codons, Fop)、亲水性指数(grand average of hydropathicity, GRAVY)。T3s、C3s、A3s、G3s分别表示第三位碱基上T/C/A/G的频率。CAI值被广泛用于评估基因表达水平, 其数值范围在0~1, 其数值越大表明密码子偏好性越强[19]。CBI反映了一个具体基因中高表达优越密码子的组分情况[4]。Fop是指某种物种高表达基因中使用频率最高的密码子[4]。GRAVY反映蛋白质的疏水性对密码子使用偏好的影响。Aromo反映芳香族蛋白质对密码子使用偏好的影响。

1.3 同义密码子相对使用度(RSCU)和相对同义密码子使用频率(RFSC)的结果分析

RSCU是指某一特定密码子在使用频率与其无偏好性使用时预期频率之间的比值[20]。虽然RSCU值与基因长度和氨基酸丰度无关, 但它能直接反映出密码子使用的偏好性程度。如果某一密码子的RSCU值=1, 表明该密码子的使用没有偏好。而当RSCU值>1, 则表明该密码子的使用频率较高, 反之亦然[4]。RFSC是某一密码子在样本中的实际观察值与该密码子对应的氨基酸在样本中的实际观察值的比值[21]。

1.4 高频密码子(high frequency codon, HF)的筛选

依据两种茶树基因组所有密码子的RFSC结果筛选高频密码子, 筛选原则如下: 某一个密码子的RFSC>60%; 或某一个密码子的RFSC超过其对应氨基酸的同义密码子平均频率的0.5倍[21]。

1.5 密码子使用频率(frequency)的比较分析

密码子使用频率(某密码子个数占该生物编码基因总密码子个数的千分比)的比值是衡量生物间密码子使用偏好性的指标之一, 当该比值≥2或≤0.5时, 表示两种生物的密码子偏好性差异较大, 反之表示偏好性差异较小[3]。为了充分了解两种茶树的密码子使用模式, 从Codon Usage Database(http://www.kazusa.or.jp/codon/)下载双子叶植物拟南芥(*Arabidopsis thaliana*; http://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=3702)、烟草(*Nicotiana tabacum*; http://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=4097)、毛果杨(*Populus trichocarpa*; http://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=3694)和单子叶植物水稻(*Oryza sativa*; http://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=311553)以及原核模式生物大肠杆菌(*Escherichia coli*; http://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=199310)、真核模式生物酿酒酵母(*Saccharomyces cerevisiae*; http://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=4932)的密码子使用偏好性数据, 将两种茶树的密码子使用频率与前述6种生物(拟南芥、毛果杨、烟草、水稻、大肠杆菌和酿酒酵母)的密码子使用频率进行比较。

1.6 有效密码子数(effective number of codons, ENc)-GC3s绘图

ENc被用于描述密码子使用偏离随机选择的程度, 能反映密码子家族中同义密码子非均衡使用的偏好程度[22]。GC3s表示第3位密码子G+C含量在基因碱基总量中所占的比率。ENc-plot是以ENc值为纵坐标, GC3s值为横坐标进行的作图分析。该方法用于分析各基因密码子的使用特征, 并用于探究基因碱基组成和密码子偏好性之间的关系[22]。有效密码子数ENc的数值范围为20~61, “ENc=20”指每个氨基酸只使用一个同义密码子, “ENc=61”指每个同义密码子被平均使用, ENc值越小表明密码子偏好性越强[23], 而且当ENc值≤35时, 基因密码子具有显著的使用偏好性[23]。当基因的密码子使用仅受G+C突变偏倚的约束时, 预测值在预期曲线上或恰好低于预期曲线[22], 当密码子使用受密码子优化选择时, ENc值远低于预期曲线。

1.7 PR2-plot

PR2-plot分析是以G3/(G3+C3)作为横坐标, A3/(A3+T3)作为纵坐标, 对氨基酸进行密码子第3位上4种碱基组成情况的分析[24]。由中心点(A=T、C=G)发出的矢量表示该碱基偏移的程度和方向。

1.8 对应分析

使用CodonW 1.4.2中的对应分析来研究基因组中的密码子使用变异情况[25]。此外, 利用SPSS 22.0软件计算出两种茶树基因组轴1(Axis 1)和密码子使用指数的相关性指数。氨基酸共有59种密码子(不包括Met、Trp和3种终止密码子), 所以根据基因在59维空间中的同义密码子使用情况来绘制基因的分布, 即每个基因对应59个坐标。轴1是捕获密码子使用的大部分变化的轴, 随后的每个轴解释方差的减少量。因此, 通过59维空间中这些轴的主要趋势可以得知这些轴占基因变异的最大分数, 从而分析密码子使用变异的主要来源。分析两种茶树基因在轴1的

位置与14个密码子使用指数的相关系数, 14个密码子使用指数的相关内容如“1.2密码子使用指数”所述。

2 结果

2.1 密码子偏好性比较

2.1.1 密码子的碱基组成分析 分别将两种茶树基因组中筛选过后的CDS序列作为一个整体, 运用Python语言计算密码子GC1、GC2、GC3和3个位置平均GC百分率。如表1所示, 两种茶树密码子3个位置平均GC%结果相近, 但GC1%、GC2%、GC3%的碱基组成却存在一定的差异; GC1%、GC2%、GC3%及3个位置平均GC%均<50%, 表明两种茶树基因组在编码蛋白的密码子使用上都倾向于A/T。

2.1.2 有效密码子数(ENc)分析 ENc值不仅能反映基因在编码蛋白质时对密码子的选择性强弱[22], 还可以确定内源基因表达量的相对高低[4]; ENc值的变化范围为20~61, 当数值越接近20, 密码子的偏好性越强, 相应内源基因的表达量越高, 反之亦然[4]。如表2所示, 两种茶树很少一部分基因在密码子使用上有很强的偏好性(CSA: 0.18%, CSS: 0.11%), 也有少部分基因的密码子使用无偏好性(CSA: 2.82%, CSS: 2.00%), 而绝大多数基因对密码子的使用有偏好性(CSA: 97.00%, CSS: 97.89%), 说明针对两种茶树的全基因组数据开展密码子偏好性研究显得尤为重要。

2.1.3 同义密码子相对使用度(RSCU)和相对同义密码子使用频率(RFSC)的结果分析 从两种茶树的RSCU值(表3)可以看出, 在两种茶树的密码子中存在偏好性的密码子为: CSA有28个, CSS有26个; 其中, 相对应的以A/T结尾的密码子分别有24和23个, 各占总偏好性密码子的85.71%和88.46%。两种茶树有25个相同的偏好密码子, 其中以A/T结尾的有22个, 占相同偏好密码子的88.00%。结果说明, 两种茶树基因组中偏好密码子更倾向于使用以A/T结尾的密码子。

两种茶树基因组密码子的RSCU变化范围分别为0.38~1.84和0.43~2.14, 而且最明显的是两种茶树的最高RSCU值均属于编码Arg的AGA; 另外, 两种茶树编码Arg的AGG的RSCU值均>1, 说明它们在编码Arg方面具有极其相似的偏好性。除此之外, CSA的最低RSCU值属于编码Ala的GCG, 而CSS的最低RSCU值属于编码Arg的CGC。

2.1.4 高频密码子(HF)的筛选 由表4可以明显看出, 两种茶树的高频密码子具有很强的共性, 但也存在一些差异。具体表现为: 两种茶树基因组中相同的高频密码子有5个, 分别为AGA、AGG、TTG、GAT、TGA。不同之处在于, CSA的GCT、GTT、CAT、AAT为高频密码子, CSS的CAA是高频密码子。在其他的RFSC较高的同义密码子中, 两种茶树都有GGA、CCA、ATT、TGT、GAA、TTT、TAT、ATG、TGG。

表1 两种茶树全基因组CDS序列的不同位置GC含量

Table 1 GC content at different locations in CDS sequences of two species of *Camellia sinensis* genome

GC含量 GC content	“云抗十号” CSA	“舒茶早” CSS
GC1%	22.58%	45.16%
GC2%	40.00%	30.00%
GC3%	40.00%	23.33%
The average of GC% at three locations	34.19%	32.83%

表2 基于ENc的两种茶树全基因组CDS分布

Table 2 CDS distributions of two species of *Camellia sinensis* genome based on ENc

组别 Groups	“云抗十号”的个数(%) Number of CSA (%)	“舒茶早”的个数(%) Number of CSS (%)
≤35	55 (0.18%)	38 (0.11%)
36-60	30 220 (97.00%)	33 200 (97.89%)
61	879 (2.82%)	679 (2.00%)
Total	31 154 (100%)	33 917 (100%)

表3 两种茶树基因组密码子的RSCU和RSFC值

Table 3 The RSCU and RSFC values of codon in two species of *Camellia sinensis* genome

氨基酸 Amino acid	密码子 Codon	“云抗十号” CSA		“舒茶早” CSS	
		同义密码子相对使用度 RSCU	相对同义密码子使用频率 RFSC	同义密码子相对使用度 RSCU	相对同义密码子使用频率 RFSC
A (Ala)	GCA	**1.31**	32.64	**1.35**	33.72
A	GCC	0.78	19.61	0.86	21.53
A	GCG	0.38	9.46	0.45	11.13
A	GCT	**1.53**	38.29	**1.34**	33.61
C (Cys)	TGC	0.89	44.59	0.97	48.40
C	TGT	**1.11**	55.41	**1.03**	51.60
D (Asp)	GAC	0.64	31.79	0.69	34.38
D	GAT	**1.36**	68.21	**1.31**	65.62
E (Glu)	GAA	**1.05**	52.67	**1.13**	56.63
E	GAG	0.95	47.33	0.87	43.37
F (Phe)	TTC	0.86	43.17	0.94	47.19
F	TTT	**1.14**	56.83	**1.06**	52.81
G (Gly)	GGA	**1.23**	30.83	**1.37**	34.34
G	GGC	0.72	17.87	0.78	19.41
G	GGG	0.87	21.67	0.85	21.27
G	GGT	**1.18**	29.54	1.00	24.98
H (His)	CAC	0.76	38.12	0.82	40.76
H	CAT	**1.24**	61.88	**1.18**	59.24
I (Ile)	ATA	0.74	24.50	0.75	24.95
I	ATC	0.83	27.78	0.97	32.36
I	ATT	**1.43**	47.72	**1.28**	42.69
K (Lys)	AAA	0.95	47.75	**1.01**	50.63
K	AAG	**1.05**	52.25	0.99	49.37
L (Leu)	CTA	0.62	10.34	0.64	10.61
L	CTC	0.91	15.13	0.95	15.86
L	CTG	0.76	12.65	0.94	12.26
L	CTT	**1.38**	22.95	**1.18**	19.62
L	TTA	0.74	12.29	0.73	12.23
L	TTG	**1.60**	26.64	**1.56**	26.07
M (Met)	ATG	1.00	100.00	1.00	100.00
N (Asn)	AAC	0.77	38.68	0.84	41.96
N	AAT	**1.23**	61.32	**1.16**	58.04
P (Pro)	CCA	**1.44**	36.01	**1.47**	36.83
P	CCC	0.68	16.94	0.79	19.73
P	CCG	0.50	12.38	0.58	14.60
P	CCT	**1.39**	34.67	**1.15**	28.83
Q (Gln)	CAA	**1.15**	57.71	**1.22**	61.23
Q	CAG	0.95	42.29	0.87	38.77
R (Arg)	AGA	**1.84**	30.69	**2.14**	35.61
R	AGG	**1.56**	25.98	**1.50**	25.03
R	CGA	0.76	12.68	0.81	13.45
R	CGC	0.51	8.53	0.43	7.19
R	CGG	0.63	10.45	0.62	10.32

(续表3)

氨基酸 Amino acid	密码子 Codon	“云抗十号” CSA		“舒茶早” CSS	
		同义密码子相对使用度 RSCU	相对同义密码子使用频率 RFSC	同义密码子相对使用度 RSCU	相对同义密码子使用频率 RFSC
R	CGT	0.70	11.67	0.50	8.40
S (Ser)	AGC	0.79	13.24	0.92	15.41
S	AGT	**1.01**	16.91	0.96	15.99
S	TCA	**1.34**	22.41	**1.36**	22.62
S	TCC	0.83	13.85	0.91	15.25
S	TCG	0.52	8.70	0.57	9.46
S	TCT	**1.49**	24.89	**1.28**	21.27
T (Thr)	ACA	**1.32**	32.91	**1.40**	35.02
T	ACC	0.90	22.58	0.96	24.08
T	ACG	0.40	10.07	0.47	11.79
T	ACT	**1.38**	34.43	**1.16**	29.11
V (Val)	GTA	0.61	15.37	0.65	16.31
V	GTC	0.70	17.57	0.77	19.25
V	GTG	**1.15**	28.82	**1.20**	29.97
V	GTT	**1.53**	38.24	**1.38**	34.47
W (Trp)	TGG	1.00	100.00	1.00	100.00
Y (Tyr)	TAC	0.83	41.52	0.85	42.40
Y	TAT	**1.17**	58.48	**1.15**	57.60
*	TAA	0.81	27.03	0.74	24.74
*	TAG	0.67	22.17	0.63	21.14
*	TGA	**1.52**	50.81	**1.62**	54.11

加粗的数字表示偏好密码子(RSCU>1), “*”表示终止子。

Bold numbers indicate preference codons (RSCU>1). The asterisk indicates the stop codon.

2.1.5 两种茶树与其他6种生物的密码子使用频率比较 由表5密码子使用频率比较结果可知, CSA与6种生物密码子使用偏差较大的密码子分别有3个、3个、3个、10个、23个和9个, 表明CSA与拟南芥、毛果杨和烟草的密码子使用频率差异最小(都仅3个), 而与大肠杆菌的密码子使用频率差别最大(23个); CSA与6种生物在TAG和TGA两个密码子的使用上存在较大差异; CSS与6种生物密码子使用偏差较大的密码子分别有5个、4个、5个、14个、24个和12个, 揭示CSS与毛果杨的密码子使用频率差别最小(仅4个), 其次与拟南芥和烟草的密码子使用频率差异较小(都仅5个), 而与大肠杆菌的密码子使用频率差别最大(24个); CSS与6种生物在GAG、TAA、TAG、TGA这4个密码子的使用上差异较大。此外, CSA与CSS的基因组密码子使用频率具有很强的相似性, 但也存在一定差异, 尤其是CCG、CCT、TCC、ACA这4个密码子的使用频率差异较大。CSA中, TCC密码子使用频率是CSS中TCC密码子使用频率的13.6倍; 而CSA中, CCG、CCT、ACA这3个密码子的使用频率相对CSS则较低, 两种茶树中3个密码子的使用频率比值分别为0.23、0.23、0.20。如果为CSA选择基因异源表达受体系统时优先考虑拟南芥、毛果杨和烟草; 那么为CSS选择基因异源表达受体系统时则最优选择毛果杨, 其次为拟南芥和烟草; 但是, 当CSA和CSS选择大肠杆菌作为基因受体系统时, 需要改造较多的密码子。

2.2 密码子使用变异来源分析

2.2.1 ENc-GC3s图分析 对两种茶树基因组密码子使用的ENc和GC3s进行绘图分析(图1)。由图1可知, 两种茶树的大部分基因的ENc值和标准曲线非常接近, 拟合程度较高, 表明它们的密码子偏好性主要是受第三位密码子(GC3s)的核苷酸组成差异的影响, 即受突变压力的影响; 但也有部分基因位于标准曲线的外侧或下方, 由此说明, 突变压力和自然选

表4 两种茶树基因组的高频密码子

Table 4 High frequency codon of two species of *Camellia sinensis* genome

氨基酸 Amino acid	“云抗十号” CSA	“舒茶早” CSS
R (Arg)	AGA^{h}, AGG^{h}	AGA^{h}, AGG^{h}
L (Leu)	TTG^{h}	TTG^{h}
S (Ser)	TCT	TCA
A (Ala)	GCT^{h}	GCA
G (Gly)	GGA	GGA
P (Pro)	CCA	CCA
T (Thr)	ACT	ACA
V (Val)	GTT^{h}	GTT
I (Ile)	ATT	ATT
C (Cys)	TGT	TGT
D (Asp)	GAT^{h}	GAT^{h}
E (Glu)	GAA	GAA
F (Phe)	TTT	TTT
H (His)	CAT^{h}	CAT
K (Lys)	AAG	AAA
N (Asn)	AAT^{h}	AAT
Q (Gln)	CAA	CAA^{h}
Y (Tyr)	TAT	TAT
M (Met)	ATG	ATG
W (Trp)	TGG	TGG
STOP	TGA^{h}	TGA^{h}

“h”表示高频密码子, 其余为RFSC较高的同义密码子。

The “h” superscript indicates the high frequency codon, the rest of codon is synonymous codon with higher RFSC.

表5 两种茶树密码子使用频率对比、两种茶树分别与6种生物密码子使用频率对比

Table 5 The usage frequency of two species of *Camellia sinensis* codon was compared, and the usage frequency of two species of *Camellia sinensis* codon were compared with six other kinds of biological codon respectively

密码子 Codon	CSA/A	CSA/P	CSA/N	CSA/O	CSA/E	CSA/S	CSS/A	CSS/P	CSS/N	CSS/O	CSS/E	CSS/S	CSA/ CSS
GCA	1.18	1.03	0.89	1.19	1.00	1.27	0.92	0.80	0.70	0.93	0.78	1.00	1.03
GCC	1.20	1.27	0.99	0.40	0.49	0.98	1.00	1.05	0.82	0.33	0.41	0.82	0.87
GCG	0.67	1.62	1.03	0.23	0.19	0.97	0.59	1.44	0.92	0.20	0.17	0.86	0.97
GCT	0.86	1.10	0.78	1.24	1.55	1.14	0.57	0.73	0.52	0.82	1.03	0.76	0.99
TGC	1.34	1.08	1.34	0.78	1.39	2.00	2.29	1.85	2.29	1.33	2.39	3.43	1.13
TGT	1.14	1.07	1.22	1.93	2.17	1.48	1.67	1.57	1.79	2.83	3.19	2.17	0.92
GAC	0.90	1.09	0.92	0.55	0.84	0.77	0.64	0.77	0.66	0.39	0.60	0.55	0.94
GAT	0.91	0.80	0.90	1.32	1.04	0.89	0.58	0.51	0.57	0.84	0.66	0.56	0.78
GAA	0.93	0.78	0.88	1.47	0.83	0.70	0.76	0.64	0.72	1.20	0.68	0.57	1.18
GAG	0.89	0.88	0.97	0.74	1.61	1.49	0.07	0.06	0.07	0.05	0.12	0.11	0.91
TTC	0.88	1.04	1.01	0.82	1.08	0.99	1.01	1.20	1.16	0.93	1.24	1.14	1.04
TTT	1.10	0.93	0.96	1.84	1.04	0.92	1.07	0.90	0.93	1.79	1.01	0.90	1.05
GGA	0.84	0.90	0.88	1.28	2.26	1.87	0.86	0.92	0.90	1.31	2.32	1.92	1.34
GGC	1.29	1.16	1.06	0.40	0.42	1.21	1.28	1.16	1.05	0.40	0.42	1.21	1.10
GGG	1.40	1.24	1.36	0.84	1.26	2.38	1.27	1.13	1.23	0.76	1.15	2.16	1.14
GGT	0.88	1.08	0.87	1.32	0.80	0.82	0.69	0.85	0.68	1.03	0.62	0.64	1.16
CAC	1.12	1.17	1.12	0.70	0.99	1.25	1.39	1.46	1.39	0.88	1.23	1.55	1.11
CAT	1.14	0.99	1.18	1.40	1.16	1.16	1.27	1.10	1.31	1.55	1.29	1.29	0.86
ATA	1.04	0.87	0.93	1.48	2.42	0.73	1.00	0.84	0.90	1.43	2.33	0.71	0.94

(续表5)

密码子 Codon	CSA/A	CSA/P	CSA/N	CSA/O	CSA/E	CSA/S	CSS/A	CSS/P	CSS/N	CSS/O	CSS/E	CSS/S	CSA/CSS
ATC	0.80	0.97	1.07	0.76	0.61	0.86	0.88	1.07	1.17	0.84	0.67	0.95	0.87
ATT	1.18	0.87	0.92	1.79	0.85	0.85	1.00	0.73	0.77	1.51	0.72	0.71	1.27
AAA	0.90	0.81	0.85	1.73	0.84	0.66	0.84	0.76	0.80	1.62	0.78	0.62	0.90
AAG	0.93	0.93	0.91	0.94	2.84	0.99	0.77	0.78	0.75	0.78	2.36	0.82	1.03
CTA	1.03	0.84	1.09	1.33	2.55	0.76	1.09	0.90	1.15	1.41	2.71	0.81	0.89
CTC	0.93	1.06	1.21	0.58	1.36	2.77	1.01	1.15	1.32	0.63	1.47	3.00	1.19
CTG	1.27	0.85	1.22	0.59	0.25	1.19	1.63	1.09	1.56	0.76	0.31	1.52	0.95
CTT	0.94	0.78	0.94	1.49	1.94	1.84	0.83	0.69	0.84	1.32	1.71	1.63	0.95
TTA	0.96	0.81	0.91	1.99	0.87	0.46	0.98	0.84	0.93	2.05	0.90	0.48	0.86
TTG	1.26	1.03	1.18	1.79	1.88	0.97	1.27	1.04	1.19	1.81	1.90	0.98	1.51
ATG	1.02	1.07	1.00	1.05	0.93	1.20	0.98	1.02	0.96	1.00	0.89	1.14	1.20
AAC	0.79	1.06	0.92	0.89	0.77	0.66	0.74	0.99	0.86	0.83	0.72	0.62	1.28
AAT	1.17	0.94	0.93	1.73	1.40	0.73	0.96	0.77	0.76	1.41	1.15	0.60	1.13
CCA	1.07	1.03	0.87	1.21	2.02	0.94	1.03	1.00	0.84	1.17	1.96	0.91	1.14
CCC	1.52	1.55	1.22	0.67	1.39	1.19	1.68	1.72	1.35	0.74	1.54	1.31	1.10
CCG	0.69	1.48	1.18	0.33	0.27	1.11	0.77	1.65	1.32	0.37	0.30	1.25	0.23
CCT	0.88	1.03	0.88	1.22	2.27	1.23	0.70	0.82	0.70	0.96	1.79	0.97	0.23
CAA	1.11	1.03	1.04	1.59	1.43	0.79	1.30	1.21	1.22	1.87	1.69	0.93	0.90
CAG	1.04	0.90	1.05	0.76	0.53	1.30	1.05	0.91	1.07	0.77	0.54	1.32	0.80
AGA	0.89	0.87	1.06	1.70	5.86	0.80	1.26	1.23	1.50	2.40	8.28	1.13	0.85
AGG	1.31	1.13	1.18	0.90	7.57	1.56	1.53	1.33	1.38	1.06	8.88	1.83	0.98
CGA	1.11	1.28	1.32	1.10	1.80	2.34	1.44	1.65	1.71	1.42	2.33	3.02	1.23
CGC	1.24	1.05	1.21	0.29	0.22	1.82	1.28	1.08	1.24	0.30	0.23	1.87	1.07
CGG	1.18	1.01	1.56	0.43	0.92	3.40	1.42	1.22	1.88	0.52	1.10	4.09	1.07
CGT	0.72	0.87	0.86	0.90	0.32	1.01	0.63	0.77	0.76	0.79	0.28	0.89	1.20
AGC	1.03	1.03	1.16	0.73	0.73	1.19	1.26	1.26	1.43	0.89	0.89	1.46	1.58
AGT	1.06	0.98	1.12	1.69	1.56	1.05	1.06	0.98	1.11	1.68	1.56	1.04	1.40
TCA	1.08	1.00	1.12	1.59	2.52	1.05	1.15	1.06	1.19	1.69	2.69	1.12	1.22
TCC	1.09	1.42	1.19	0.75	1.37	0.86	1.26	1.64	1.39	0.87	1.59	1.00	13.60
TCG	0.82	1.53	1.44	0.62	0.88	0.89	0.94	1.75	1.65	0.71	1.01	1.02	0.68
TCT	0.87	1.07	1.09	1.72	2.51	0.93	0.78	0.96	0.99	1.55	2.27	0.84	0.58
ACA	1.03	1.06	0.93	1.40	1.98	0.91	1.09	1.12	0.98	1.47	2.08	0.96	0.20
ACC	1.08	1.34	1.15	0.75	0.49	0.88	1.14	1.42	1.21	0.79	0.52	0.93	0.64
ACG	0.65	1.13	1.10	0.44	0.34	0.62	0.75	1.31	1.28	0.50	0.39	0.72	1.14
ACT	0.97	1.18	0.84	1.60	1.87	0.84	0.81	0.99	0.70	1.34	1.56	0.70	0.97
GTA	1.00	0.97	0.87	1.46	0.89	0.84	0.88	0.86	0.77	1.28	0.79	0.74	0.77
GTC	0.89	1.00	1.02	0.56	0.75	0.96	0.80	0.91	0.93	0.51	0.68	0.87	0.83
GTG	1.07	1.10	1.11	0.77	0.73	1.72	0.92	0.95	0.96	0.66	0.63	1.49	1.00
GTT	0.91	1.02	0.92	1.59	1.34	1.12	0.68	0.76	0.69	1.19	1.00	0.83	0.81
TGG	1.24	1.11	1.27	1.12	1.02	1.48	1.92	1.73	1.97	1.74	1.58	2.31	0.71
TAC	0.84	1.21	0.85	0.76	0.95	0.77	0.76	1.10	0.77	0.69	0.86	0.70	0.85
TAT	1.11	0.99	0.91	1.61	0.98	0.86	0.97	0.87	0.79	1.42	0.86	0.75	1.28
TAA	2.79	6.28	2.28	3.59	1.26	2.28	11.90	26.80	9.74	15.30	5.36	9.74	1.00
TAG	4.12	5.15	4.12	2.58	6.87	4.12	18.30	22.90	18.30	11.40	30.50	18.30	0.97
TGA	3.93	6.73	4.71	3.93	4.28	6.73	19.50	33.50	23.40	19.50	21.30	33.50	1.10

A: 拟南芥(*Arabidopsis thaliana*); P: 毛果杨(*Populus trichocarpa*); N: 烟草(*Nicotiana tabacum*); O: 水稻(*Oryza sativa*); E: 大肠杆菌(*Escherichia coli*); S: 酿酒酵母(*Saccharomyces cerevisiae*); 6种生物的密码子使用频率数据均来源于Codon Usage Database(http://www.kazusa.or.jp/codon/), 详见前文“1.5密码子使用频率(Frequency)的比较分析”所述; 外侧有框线的数字表示密码子使用偏差较大。

A: *Arabidopsis thaliana*; P: *Populus trichocarpa*; N: *Nicotiana tabacum*; O: *Oryza sativa*; E: *Escherichia coli*; S: *Saccharomyces cerevisiae*. The codon usage frequency data of the 6 species are derived from Codon Usage Database. The details are described in the preceding “1.5 comparative analysis of codon usage frequency”. The number of the outer frame lines indicates that the codon usage is quite different.

择等其他因素也共同参与了两种茶树基因组密码子使用偏好性的形成。

2.2.2 PR2-plot分析 通常认为, 当核苷酸碱基变异发生在密码子第三位上, 即基因密码子偏好性完全受突变影响时, 基因或基因组中简并密码子的A、T和C、G所占比例相近[24], 但在对两种茶树基因组进行PR2-plot分析时发现, GC和AT使用频率是不平衡的(图2), 说明两种茶树基因组的碱基组成不仅受碱基突变的影响, 还受到自然选择的影响。除了这一共性, 两种茶树还存在如下差异: CSA(图2A)中, A和T使用主要分布在0.25~0.67的范围, G和C使用主要分布在0.21~0.83的范围; 而在CSS(图2B)中, A和T主要分布在0.24~0.80的范围, G和C主要分布在0.27~0.62的范围。以上表明, CSS较CSA在碱基C的使用上存在更强的偏好性。

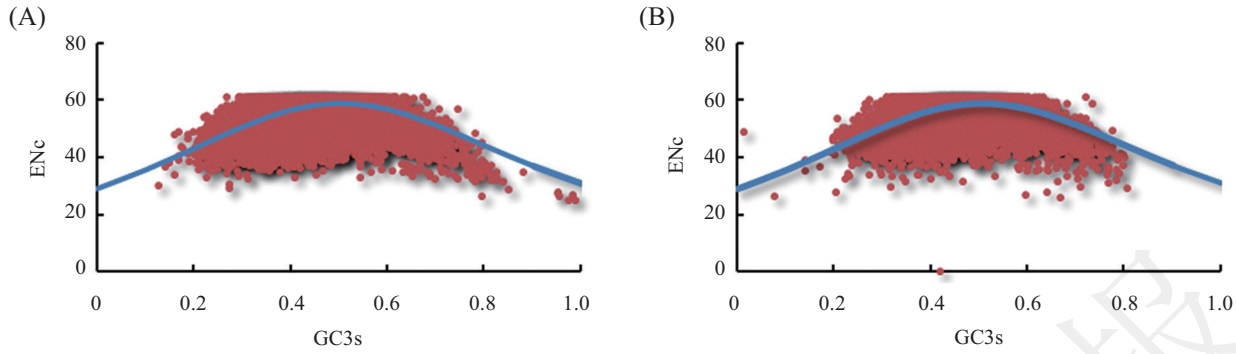
2.2.3 对应分析 本研究通过基于RSCU的对应分析来揭示影响两种茶树基因组密码子使用模式的主要因素。对应分析用于比较59个密码子的使用模式, 其结果产生一系列正交轴, 反映了密码子使用中变异的趋势。图3中每个点代表一个与从对应分析获得的轴1和轴2上的坐标相对应的基因, 用不同的颜色标记不同GC含量的基因, 蓝色表示GC%<45%, 棕色表示45%≤GC%<60%, 绿色表示GC%≥60%, 旨在研究GC含量对密码子使用偏倚的影响。如图3结果所示, CSA、CSS的G、C含量>60%的基因极少, 而且G、C含量<45%以及介于45%~60%之间的基因无明显分离。为了确定导致基因沿轴1和轴2分散的因素, 计算两种茶树基因在轴1的位置与14个密码子使用指数的相关系数(表6)。由表6结果可知, 大部分指数与轴1呈极显著相关, 说明很多因素对密码子偏好性有明显的影响; 对比两种植物发现CSA的轴1与14种密码子使用指数全部呈极显著相关, CSS个别指数无显著相关。CSA和CSS的轴1与GC3s呈极显著相关(P≤0.01), 表明突变压力中碱基组成是密码子使用偏好性的主要影响因素; 另外, 两种茶树的轴1与CAI均呈极显著相关, 说明基因表达水平也是影响密码子使用变异的因素之一; 除此之外, CSA、CSS的GC3s相关系数均>CAI, 说明碱基组成的变异比基因表达水平对密码子偏好性的影响更大; 此外, 基因长度、氨基酸的疏水性以及芳香性与密码子轴1也有显著相关性, 表明这些因素都影响到密码子使用偏好性, 但相对于碱基组成和基因表达水平, 其对密码子使用变异的影响要小得多。

3 讨论

在生物漫长的进化历程中, 每个物种都会形成自身特定的密码子用法以适应环境[26]。通过分析基因组密码子的使用变异来源以及密码子偏好模式, 确定高频密码子, 用于重建外源基因的密码子, 提高外源基因在宿主细胞中的表达效率, 这将对植物的基因工程研究和遗传进化研究具有重要的意义[4]。

基因组CDS序列分析揭示, 两种茶树基因组的编码蛋白在密码子使用上多倾向于A/T, 该结果与丁兆堂课题组关于茶树单基因的研究结果(*ICE1*[13]、*SAD*[14]、*GPAT*[15]、*SPDS*[16])一致; 据分析, 与GC3s值较高(GC3s值>50%, 表现为: 密码子使用上偏好于G/C)的单子叶植物不同, 双子叶植物的GC3s值往往<50%(密码子使用上偏好于A/T)[27]。RSCU值分析结果表明, 两种茶树基因组中存在偏好性的密码子多以A/T结尾(CSA: 85.71%, CSS: 88.46%), 该结果与茶树*ICE1*[13]、*SAD*[14]、*GPAT*[15]和*SPDS*[16]单基因密码子偏好性研究结果一致。Novembre[28]的研究结论可用于解释该研究结果: 排除自然选择影响下, 碱基突变的压力将影响到同义密码子第3位碱基的组成, 当A/T突变为G/C压力高时, 会引起密码子第3位多为G/C, 反之多为A/T。由本研究中偏好密码子多以A/T结尾可见, 两种茶树基因组中碱基"G/C的突变压力">"A/T的突变压力"。PR2-plot分析揭示, 两种茶树基因组的碱基组成不仅受到突变压力的影响, 还受自然选择的影响, 这与苹果[9]和桑树[10]基因组密码子偏好性分析研究结果一致。此外, ENc分析及其对应分析和GC3s研究均表明, 前述两种影响因素中突变压力占主导地位, 究其原因, 可能如Grantham等[26]提出的"基因组假说(Genome Hypothesis)"所言: 任何给定基因组中密码子应用上的偏好性都具有物种特异性。

两种茶树基因组拥有5个相同的高频密码子: AGA、AGG、TTG、GAT、TGA; 同时, CSA含有4个特有的高频密码子: GCT、GTT、CAT、AAT, CSS仅有CAA 1个特有的高频密码子。这些高频密码子研究结果不仅有利于密码子的优化改造, 还有益于进一步分析茶树基因表达与密码子偏好性之间的关系。对密码子使用频率进行比较时发现, CSA

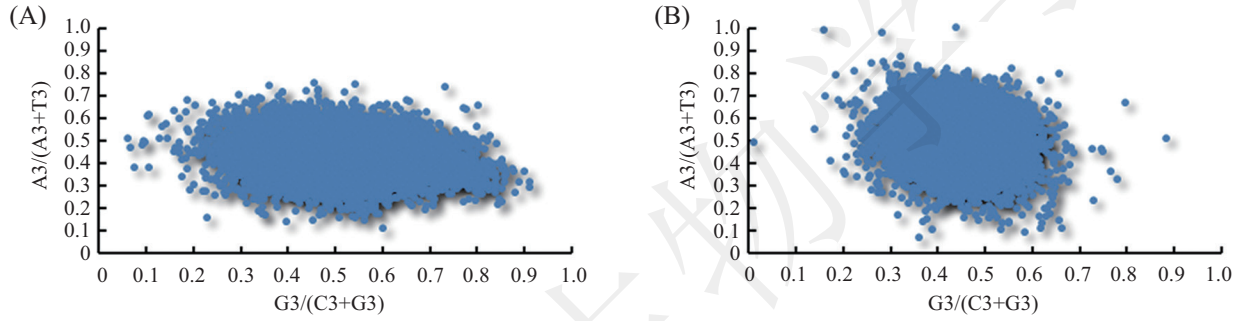


A: CSA全基因组ENc-GC3s分析结果; B: CSS全基因组ENc-GC3s分析结果。

A: the ENc-GC3s analysis results of CSA genome; B: the ENc-GC3s analysis results of CSS genome.

图1 两种茶树全基因组ENc-GC3s绘图

Fig.1 ENc-GC3s plot of two species of *Camellia sinensis* genome

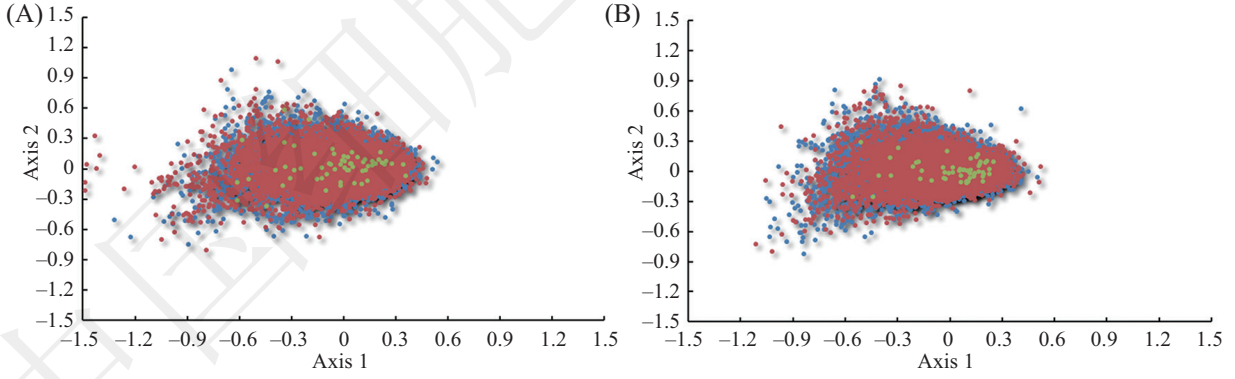


A: CSA全基因组PR2-plot分析; B: CSS全基因组PR2-plot分析。

A: the PR2-plot analysis results of CSA genome; B: the PR2-plot analysis results of CSS genome.

图2 两种茶树基因组PR2-plot分析

Fig.2 PR2-plot analysis of two species of *Camellia sinensis* genome



A: CSA全基因组基于RSCU值的对应分析图; B: CSS全基因组基于RSCU值的对应分析图。

A: the corresponding analysis results of CSA genome based on RSCU value; B: the corresponding analysis results of CSS genome based on RSCU value.

图3 两种茶树全基因组基于RSCU值的对应分析图

Fig.3 The corresponding analysis diagram of two species of *Camellia sinensis* genome based on RSCU values

基因的异源表达受体系统优先选择拟南芥、毛果杨和烟草; CSS基因的异源表达受体系统首选毛果杨, 其次选择拟南芥和烟草; 但是, 当CSA和CSS的基因在大肠杆菌中异源表达时, 需要改造的密码子较多。比较密码子使用频率时, 研究揭示两种茶树基因组的密码子使用频率基本相似, 仅CCG、CCT、TCC、ACA 4个密码子的使用频率差别较大, 尤其CSA中TCC密码子使用频率是CSS中TCC密码子使用频率的13.6倍, CSA中CCG、CCT、ACA这3个密码子的使用频率相对CSS则较低。

本文首次基于两种茶树全基因组水平开展密码子偏好性的比较研究, 分析结果有益于比较分析

表6 两种茶树基因组轴1(Axis 1)和密码子使用指数的相关性分析

Table 6 Correlation analysis of Axis 1 and codon usage index of two species of *Camellia sinensis* genome

密码子使用指数 Codon usage index	“云抗十号” CSA	“舒茶早” CSS
T3s	0.749**	0.790**
C3s	–0.738**	–0.758**
A3s	–0.503**	–0.522**
G3s	–0.465**	–0.462**
CAI	–0.078**	–0.072**
CBI	0.640**	0.616**
Fop	0.966**	0.963**
ENc	0.086**	0.112**
GC3s	0.169**	0.181**
GC	0.787**	0.780**
L_sym	–0.112**	–0.206**
L_aa	1.000**	1.000**
GRAVY	–0.042**	–0.012*
Aromo	0.354**	0.380**

表6中英文缩写详见前文“1.2密码子使用指数”; *P≤0.05, **P≤0.01。

The abbreviations in Table 6 are described in the preceding “1.2 codon usage index”; *P≤0.05, **P≤0.01.

两种茶树的遗传进化规律、为茶树选择合适的基因异源表达受体系统及优化出适合茶树基因表达的密码子成分, 最终实现构建稳定高效的茶树基因表达系统的研究目标。随着第三代基因编辑技术CRISPR/Cas9的快速发展及应用, 期待能在本研究的基础上, 通过优化茶树密码子来改造用于茶树基因组编辑的Cas9基因, 实现提高Cas9基因在茶树中的表达[29]。

致谢——

感谢高立志团队和宛晓春团队分别攻克了“云抗10号”和“舒茶早”的全基因组测序工作, 两种茶树的全基因测序结果即为本文研究的源数据。

参考文献 (References)

1 Novoa EM, Pavon-Eternod M, Pan T, Ribas de Pouplana L. A role for tRNA modifications in genome structure and codon usage. Cell 2012; 149(1): 202-13.

2 Ikemura T. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 1985; 2(1): 13-34.

3 郭秀丽, 王玉, 杨路成, 丁兆堂. 茶树*CBF1*基因密码子使用特性分析. 遗传[Guo Xiuli, Wang Yu, Yang Lucheng, Ding Zhaotang. Analysis of codon use features of *CBF1* gene in *Camellia sinensis*. Hereditas (Beijing)] 2012; 34(12): 1614-23.

4 吴宪明, 吴松锋, 任大明, 朱云平, 贺福初. 密码子偏性的分析方法及相关研究进展. 遗传[Wu Xianming, Wu Songfeng, Ren Daming, Zhu Yunping, He Fuchu. The analysis method and progress in the study of codon bias. Hereditas (Beijing)] 2007; 29(4): 420-6.

5 Park S, Pack SP, Lee J. Expression of codon-optmized phosphoenolpyruvate carboxylase gene from *Glaciecola* sp. *HTCC2999* in *Escherichia coli* and its application for C_4 chemical production. Appl Biochem Biotechnol 2012; 167(7): 1845-53.

6 Spatz SJ, Volkening JD, Mullis R, Li F, Mercado J, Zsak L. Expression of chicken parvovirus VP2 in chicken embryo fibroblasts requires codon optimization for production of naked DNA and vectored meleagrid herpesvirus type 1 vaccines. Virus Genes 2013; 47(2): 259-67.

7 Zelasko S, Palaria A, Das A. Optimizations to achieve high-level expression of cytochrome P450 proteins using *Escherichia coli* expression systems. Protein Expr Purif 2013; 92(1): 77-87.

8 De La Torre AR, Lin YC, Van de Peer Y, Ingvarsson PK. Genome-wide analysis reveals diverged patterns of codon bias, gene expression, and rates of sequence evolution in *Picea* gene families. Genome Biol Evol 2015; 7(4): 1002-15.

9 Li N, Sun MH, Jiang ZS, Shu HR, Zhang SZ. Genome-wide analysis of the synonymous codon usage patterns in apple. J Interg Agr 2016; 15(5): 983-91.

10 Wen Y, Zou ZL, Li HS, Xiang ZH, He NJ. Analysis of codon usage patterns in *Morus notabilis* based on genome and transcriptome data. Genome 2017; 60(6): 473-84.

11 Xia EH, Zhang HB, Sheng J, Li K, Zhang QJ, Kim C, *et al*. The tea tree genome provides insights into tea flavor and independent evolution of caffeine biosynthesis. Mol Plant 2017; 10(6): 866-77.

12 Wei C, Yang H, Wang S, Zhao J, Liu C, Gao L, *et al*. Draft

genome sequence of *Camellia sinensis* var. *sinensis* provides insights into the evolution of the tea genome and tea quality. Proc Natl Acad Sci USA 2018; 115(18): 4151-8.

13 时慧, 王玉, 杨路成, 丁兆堂. 茶树抗寒调控转录因子*ICE1*密码子偏性分析. 园艺学报(Shi Hui, Wang Yu, Yang Lucheng, Ding Zhaotang. Analysis of codon bias of the cold regulated transcription factor ICE1 in tea plant. Acta Horticulturae Sinica) 2012; 39(7): 1341-52.

14 Pan LL, Wang Y, Hu JH, Ding ZT, Li C. Analysis of codon use features of stearoyl-acyl carrier protein desaturase gene in *Camellia sinensis*. J Theor Biol 2013; 334(2013) : 80-6.

15 Chen L, Pan LL, Wang Y, Wang J, Ding ZT. Codon bias of the gene for chloroplast glycerol-3-phosphate acyltransferase in *Camellia sinensis* (L.) O. Kuntze. Biochem Syst Ecol 2014; 55(2014) : 212-8.

16 You E, Wang Y, Ding ZT, Zhang XF, Pan LL, Zheng C. Codon usage bias analysis for the spermidine synthase gene from *Camellia sinensis* (L.) O. Kuntze. Genet Mol Res 2015; 14(3): 7368-76.

17 柏锡, 徐建震, 李杰, 郭政, 李琳, 朱延明. 马铃薯密码子用法分析及其在t-PA基因密码子改造上的应用. 遗传[Bai Xi, Xu Jianzhen, Li Jie, Guo Zheng, Li Lin, Zhu Yanming. Analysis of codon usage in potato and its application in the modification of t-PA gene. Hereditas (Beijing)] 2004; 26(1): 75-83.

18 McInerney JO. GCUA: general codon usage analysis. Bioinformatics 1998; 14(4): 372-3.

19 Sharp PM, Li WH. An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol 1986; 24(1/2): 28-38.

20 Sharp PM, Li WH. The codon Adaptation Index－a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 1987; 15(3): 1281-95.

21 Wright F. The ‘effective number of codons’ used in a gene. Gene 1990; 87(1): 23-9.

22 林涛, 倪志华, 沈明山, 陈亮. 高频密码子分析法及其在烟草密码子分析中的应用. 厦门大学学报(自然科学版)[Lin Tao, Ni Zhihua, Shen Mingshan, Chen Liang. High-frequency codon analysis and its application in codon analysis of tobacco. Journal of Xiamen University (Natural Science)] 2002; 41(5): 551-4.

23 Ikemura T. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J Mol Biol 1981; 151(3): 389-409.

24 Sueoka N. Translation-coupled violation of Parity Rule 2 in human genes is not the cause of heterogeneity of the DNA G+C content of third codon position. Gene 1999; 238(1): 53-8.

25 Shields DC, Sharp PM. Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. Nucleic Acids Res 1987; 15(19): 8023-40.

26 Grantham R, Gautier C, Gouy M, Mercier R, Pavé A. Codon catalog usage and the genome hypothesis. Nucleic Acids Res 1980; 8(1): 49-62.

27 Murray EE, Lotzer J, Eberle M. Codon usage in plant genes. Nucleic Acids Res 1989; 17(2): 477-98.

28 Novembre JA. Accounting for background nucleotide composition when measuring codon usage bias. Mol Biol Evol 2002; 19(8): 1390-4.

29 Ma X, Zhu Q, Chen Y, Liu YG. CRISPR/Cas9 platforms for genome editing in plants: developments and applications. Mol Plant 2016; 9(7): 961-74.