Tree Genetics and Molecular Breeding 2021, Vol.11, No.1, 1-11
http://genbreedpublisher.com/index.php/tgmb

Research Article Open Access

Analysis of Codon Usage Bias in the *Platycarya* Chloroplast Genome

Wang Xiaoshuang, Wang Yanqing, Li Shihong, Liu Yingliang ✉, Zhu Bin
School of Life Science, Guizhou Normal University, Guiyang 550025, China
✉ Corresponding author email: liuyl-23@126.com

Tree Genetics and Molecular Breeding, 2021, Vol.11, No.1 doi: 10.5376/tgmb.2021.11.0001
Received: 12 Aug., 2021
Accepted: 20 Aug., 2021
Published: 31 Aug., 2021

Preferred citation for this article:
Wang X.S., Wang Y.Q., Li S.H., Liu Y.L., and Zhu B., 2021, Analysis of codon usage bias in the *Platycarya* chloroplast genome, Tree Genetics and Molecular Breeding, 11(1): 1-11 (doi: 10.5376/tgmb.2021.11.0001)

Abstract To detect the codon usage characteristics of chloroplast genomes in *Platycarya* genus, the CodonW、CUSP and SPSSAU software were employed to analyze the codon usage patterns of the chloroplast genomes protein-coding sequence in *Platycarya longipes* and *Platycarya strobilacea* in this research. Results show that the GC content of chloroplast genomes was 37.75% and 37.80%, the average GC content in the 3rd position was 27.16% and 27.25%. the range of effective codon number from 35.19 to 56.98, and there were more than 2/3 genes when ENC value greater than 45, which indicated a weak preference. According to the results of neutrality plot analysis, ENC-plot analysis and PR2-plot analysis, codon bias in most *Platycarya* chloroplast genes were affected by natural selection, while a few were affected by mutations or other factors. And based on the ENC value, five groups of high-expressed and low-expressed genes were identified, 16 codons were ended with A/U among the 18 optimal codons. The research has implications on codon optimization, enhancing the expression efficiency of exogenous gene and phylogenetic analysis.

Keywords *Platycarya*; Chloroplast genomes; Codon bias; Optimal codons

Chloroplast is an important photosynthetic organelle, which provides energy for plant growth and development, the genetic information independent of nucleus and the chloroplast genome has a simple structure and a large number of gene copies (Niu et al., 2018; Wang et al., 2019). It has the characteristics of self-reproduction and maternal inheritance, and the genome size is between 120 and 160 KB (Suiura, 1992). The genetic information of the chloroplast genome is highly conserved (Wang et al., 2012), generally considered that chloroplast DNA (cpDNA) is a closed double-stranded circular molecule, and only a few linear molecules. Due to the high homology of the chloroplast gene sequence, lead to the replication rate is higher than nuclear DNA, many functions of the chloroplast genome have been annotated, thus becomes a powerful tool for plant systematics research chloroplast genome (Freitas et al., 2016; Kong and Yang, 2016). CpDNA can be divided into three categories: chloroplast photosynthetic genes, Chloroplast Expression Genes and other genes related to biosynthesis (Li et al., 2019). With the development of high-throughput sequencing technology, the chloroplast genomes of many woody plants such as *Rosa chinensis*, *Cinnamomum camphora* and *Pinus massoniana* have been sequenced, and the analysis of codon bias has been completed, reveals that the chloroplast genome is affected by natural selection in evolution, therefore, to analyze the chloroplast genome will provide inspiration for understanding evolution and natural selection mechanisms.

The genetic information carried by DNA is transmitted in the form of codon during the transformation into protein (Zhang et al., 2019). Every amino acid is encoded by 1-6 codons, the codons encoding the same amino acid are called synonymous codons. Due to gene mutation and natural selection, a species or gene tends to use one or more specific synonymous codons, which is called synonymous codon usage bias (Long et al., 2018; Wang et al, 2018; Li et al., 2019). Codon usage patterns usually represent a balance between gene mutation and neutral selection, and codon preference is also associated with phylogenetic relationships in some species (Wu, 2007; Zhao et al., 2016). Therefore, the analysis of codon usage patterns will help to explore the molecular evolution of species, the mechanism of protein action, the improvement of expression efficiency of exogenous genes and species classification (Qin et al., 2013; Qi et al., 2015).

Platycarya is a small deciduous tree of Juglandaceae, includes two species, *Platycarya longipes* and *Platycarya strobilacea*, with unique fruit morphology, which is obviously different from other genera in Juglandaceae family. The population of *Platycarya* in China is mainly distributed in Guizhou, Guangxi and Guangdong karst areas. Studies have shown that there are a large number of ellagic acid and gallic acid (Wang, 2010; Liu et al., 2016), quercetin and tannin in the fruits of *Platycarya*, and the leaves contain high levels of ascorbic acid (Yang, 2010), Polysaccharides, tannins, and flavonoids can be used in clinic. In addition, as the main constructive species in the karst ecological area, the genus *Platycarya* plays an important role in maintaining the stability of the karst forest ecosystem (Xu et al., 2019).

In this study, the base composition and codon usage patterns of chloroplast genomes of two species of *Platycarya* were analyzed, and the optimal codons were obtained, which would provide scientific reference for the application and research of chloroplast genomes and genetic engineering of *Platycarya* plants.

1 Results and Analysis

1.1 Codon composition analysis

CodonW 1.4.2 and CUSP were used to analyze the codon composition of the chloroplast genomes of P. longipes and P. strobilacea. The average GC content of the coding sequence of *P. longipes* was 37.75%, and the GC1, GC2, GC3 and GC3s contents were 46.39%, 39.61%, 27.25% and 24.11%, respectively. The GC content of *P. strobilacea* coding gene was 37.80%, GC1, GC2, GC3 and GC3s were 46.16%, 39.21%, 27.16% and 24.04%, respectively. It can be seen that the GC is not evenly distributed on the codon, the base at the third position of the codon is mainly A/T, and shows a trend of GC1 > GC2 > GC3 (Table 1).

The number of effective codons can reveal the degree of codon preference, the ENC value is negatively correlated with the degree of codon preference. An ENC value less than 45 means that there is a preference for codons. In this study, the ENC value of the chloroplast genome of *P. longipes* was 35.19~56.98, with an average of 46.37, the ENC value of the chloroplast genome of *P. strobilacea* was 35.19~53.29, with an average of 46.31. There were 74 genes in the two species with ENC values greater than 45, accounting for 70% of the total coding sequences, indicating that the codon preference of most of the protein-coding gene sequences of the chloroplast of the genus *Platycarya* is not strong. It was found that the ENC values of *rps14*, *ndhC*, *petD*, *rp12*, *ndhA* and *ndhB* were quite different between the two species, and the GC content and ENC value of most of the genes had very little difference, indicating that the chloroplast coding genes of Platycarya are relatively conserved. The differences between species are relatively small, which may be due to similar pressures on *Platycarya* during evolution.

The correlation analysis of the GC_1, GC_2, GC_{12}, GC_3, GC_{all}, GC_{3s} content, the number of codons and ENC values of the chloroplast genome in the genus *Platycarya* were performed, showed that GC_{all} was significantly correlated with GC_1, GC_2 and GC_3, GC_1 was significantly correlated with GC_2, but there was no correlation between GC_3 and GC_1, GC_2, and the three-position base composition of the codons is not similar, indicating that selection pressure plays a major role in the formation of codon preference in the chloroplast genome of *Platycarya*, the results show that ENC value is significantly correlated with GC_{3s}, indicating that the composition of the third base of synonymous codons directly affects codon preference, The correlation analysis also showed that ENC value was not only correlated with GC_{3s}, but also significantly correlated with the length of genes (Table 2; Table 3).

RSCU (Relative Synonymous Codon Usage) refers to the probability of a specific codon in a certain amino acid in synonymous codon. If the usage of codon is not preferred, the RSCU value of the codon is equal to 1, when the RSCU of an amino acid is greater than 1, it means that the usage frequency of the codon is higher. The RSCU values of Platycarya chloroplast genomes were analyzed by Codonw1.4.2, and the number of codons with RSCU>1 in the two plants was 30, except for the codon UUG encoding leucine. 29 of them all end in A or U, indicating that the codon preference of the chloroplast genome of the Platycarya ends in A and U, and UAA is preferred in the use of the stop code, the highest and lowest values of RSCU were coding UUA and CUG of leucine, the above analysis results show that the RSCU value and preference of codons between the two species maintain a high degree of consistency (Table 4).

Table 1 GC content and ENC values in different positions of codons in coding sequence of *Platycarya* cp genomes

Genes	GC content of *P. longips* (%)						GC content of *P. strobilacea* (%)					
	GC_1	GC_2	GC_3	GC_{all}	GC_{3s}	ENC	GC_1	GC_2	GC_3	GC_{all}	GC_{3s}	ENC
psbA	49.72	43.50	32.20	41.81	27.80	40.36	49.72	43.50	32.20	41.81	27.80	40.89
matK	38.51	31.43	26.72	32.22	25.00	49.65	38.58	31.50	26.57	32.22	24.80	49.87
atpA	55.12	39.76	24.80	39.9	23.20	47.44	55.31	39.76	25.00	40.03	23.40	47.50
atpF	45.40	28.83	32.52	35.58	29.50	51.28	44.32	30.81	31.35	35.50	29.40	50.69
atpI	48.79	36.69	23.39	36.29	20.90	43.79	48.79	26.29	23.79	36.29	21.30	43.83
rps2	43.04	40.51	27.43	36.99	23.70	47.81	42.62	40.51	27.43	36.85	23.70	48.21
rpoC2	45.54	36.19	26.98	36.24	25.10	48.54	45.62	36.01	27.12	36.25	25.10	48.51
rpoC1	51.89	36.55	28.99	39.15	26.60	50.28	50.73	37.32	28.28	38.78	26.10	49.53
rpoB	50.14	37.82	26.80	38.25	24.40	47.94	49.95	37.91	26.70	38.19	24.30	47.76
psbD	52.54	43.22	30.79	42.18	26.30	46.70	52.54	43.22	30.79	42.18	26.30	46.70
psbC	54.98	46.10	29.44	43.51	25.60	46.40	53.89	45.49	31.97	43.78	28.00	48.26
rps14	46.51	50.00	27.91	41.47	25.60	36.09	42.57	47.52	30.69	40.26	28.10	38.44
psaB	48.58	42.90	29.50	40.32	25.00	47.31	48.71	42.99	29.80	40.50	25.30	47.59
psaA	52.06	43.54	31.16	42.25	27.00	49.68	52.06	43.54	31.16	42.25	27.00	49.68
rps4	50.99	37.62	25.74	38.12	24.40	49.84	50.99	37.13	25.25	37.79	23.90	50.48
ndhJ	50.94	37.74	30.19	39.62	26.20	47.54	50.94	37.11	29.56	39.20	25.50	48.34
ndhK	44.93	44.49	25.55	38.33	22.50	51.10	44.93	44.49	25.55	38.33	22.50	50.8
ndhC	45.58	35.37	27.89	36.28	21.50	44.94	46.28	33.88	23.14	34.44	16.20	47.10
atpE	48.51	39.55	30.60	39.55	27.60	48.53	48.51	39.55	30.60	39.55	27.60	48.53
atpB	56.50	42.48	27.03	42.01	24.70	44.49	56.50	42.48	26.83	41.94	24.50	44.36
rbcL	58.40	43.70	28.36	43.49	25.80	46.86	58.18	44.10	29.40	43.89	26.70	47.95
accD	41.32	33.73	26.55	33.87	23.70	48.24	40.74	33.14	27.29	33.72	24.50	48.13
ycf4	42.70	40.54	31.35	38.20	27.60	50.93	42.70	40.54	31.35	38.20	27.60	50.93
cemA	37.83	28.26	32.61	32.90	29.00	39.55	37.39	28.26	33.17	33.61	28.60	39.49
petA	52.17	35.71	30.43	39.44	29.10	48.98	52.17	35.71	30.43	39.44	29.10	49.64
psbE	44.05	47.62	28.57	40.08	25.00	46.7	44.05	46.43	28.57	39.68	25.00	46.62
rps18	37.25	41.18	25.49	34.64	23.20	38.22	37.25	41.18	25.49	34.64	23.20	38.22
rp120	33.90	40.68	23.73	32.77	21.90	42.39	33.90	40.68	23.73	32.77	21.90	42.39
psbB	54.81	46.37	28.09	43.09	24.10	46.72	54.81	46.37	27.90	43.03	23.90	46.65
petB	49.08	43.56	30.67	41.10	23.60	42.96	49.07	41.67	29.17	39.97	23.10	40.78
petD	51.08	39.25	23.66	37.99	21.20	39.47	50.93	39.13	22.36	37.47	19.00	37.35
rpoA	46.20	31.00	23.71	33.64	21.60	49.01	46.50	31.61	23.40	33.84	21.20	48.41
rps11	51.80	56.12	25.18	44.36	22.60	43.24	51.80	56.12	25.90	44.60	23.30	43.24
rps8	40.74	40.74	26.67	36.05	25.20	45.34	40.74	41.48	27.41	36.54	26.00	45.91
rp114	54.47	37.40	21.95	37.94	20.20	46.48	53.66	37.40	22.76	37.94	21.00	47.13
rp116	51.67	53.33	26.67	43.89	20.70	42.63	50.00	43.68	26.47	43.38	20.60	43.11
rps3	48.23	34.51	23.45	35.40	20.40	49.98	48.64	34.09	21.82	34.85	19.30	48.89
rp122	34.13	38.92	23.95	32.34	19.70	46.32	33.33	38.67	23.33	31.78	19.10	46.21
rps19	42.11	34.74	23.16	33.33	20.90	52.36	42.11	34.74	23.16	33.33	20.90	52.36
rp12	48.51	44.03	29.85	40.80	29.00	56.98	51.27	48.00	31,64	43.46	29.90	52.68
rp123	40.43	41.49	28.72	36.88	23.30	48.65	40.43	41.49	28.72	36.88	23.30	48.65
ycf2	41.43	34.36	36.60	37.46	34.00	53.22	41.41	34.40	36.55	37.45	33.90	53.29
rps7	52.56	44.87	26.28	41.24	23.50	45.71	52.56	44.87	26.28	41.24	23.50	45.71
rps15	35.16	30.77	26.37	30.77	25.80	35.19	35.16	30.77	26.37	30.77	25.80	35.19
ndhH	50.25	35.79	26.14	37.39	21.20	46.31	50.25	35.79	26.65	37.56	21.50	47.46
ndhA	48.19	39.38	21.24	36.27	19.30	46.84	43.84	38.36	21.64	34.61	19.00	44.23
ndhI	41.57	36.75	19.88	32.73	16.50	40.11	41.57	36.75	19.28	32.53	15.80	39.84
ndhG	44.07	34.46	28.25	35.59	25.40	46.76	44.07	34.46	28.81	35.78	26.00	47.75
ndhE	41.58	33.66	19.80	31.68	16.50	42.93	41.18	33.33	21.57	32.03	17.50	43.25
psaC	45.12	53.66	26.83	41.87	22.40	46.63	45.12	53.66	26.83	41.87	22.40	46.63
ndhD	40.43	36.69	27.81	34.98	24.00	47.48	40.83	36.49	28.21	35.17	24.40	47.80
ndhF	35.36	34.43	22.30	30.69	18.10	42.93	35.71	34.91	22.51	31,04	18.50	43.14
ndhB	41.7	37.25	34.41	37.79	30.90	51.75	41.33	38.79	31.77	37.30	27.80	48.40

Table 2 Correlation analysis of GC contents and related parameters in codons of *P. longipes*

Item	GC_1	GC_2	GC_{12}	GC_3	GC_{all}	GC_{3s}	N
GC_2	0.461**						
GC_{12}	0.859**	0.851**					
GC_3	0.142	0.094	0.138				
GC_{all}	0.827**	0.805**	0.955**	0.426**			
GC_{3s}	0.108	-0.012	0.057	0.919**	0.327*		
N	-0.055	-0.27	-0.188	0.274	-0.09	0.325*	
ENC	0.171	-0.144	0.018	0.248	0.091	0.297*	0.291*

Note: **Correlations at a level of 0.01, *Correlations at a level of 0.05, N represents the number of codons

Table 3 Correlation analysis of GC contents and related parameters in codons of *P. strobilacea*

Item	GC_1	GC_2	GC_{12}	GC_3	GC_{all}	GC_{3s}	N
GC_2	0.341*						
GC_{12}	0.819**	0.819**					
GC_3	0.121	0.161	0.173				
GC_{all}	0.783**	0.753**	0.938**	0.461**			
GC_{3s}	0.115	0.116	0.141	0.928**	0.393**		
N	0.115	-0.129	-0.008	0.405**	0.102	0.426**	
ENC	0.241	-0.032	0.128	0.234	0.165	0.296*	0.361**

Note: **Correlations at a level of 0.01, *Correlations at a level of 0.05, N represents the number of codons

Table 4 The relative synonymous codon usage (RSCU) of two *Platycarya* cp genomes

Amino Acid	Codon	*P. longips*		*P. strobilacea*		Amino Acid	Codon	*P. longips*		*P. strobilacea*	
		Number	RSCU	Number	RSCU			Number	RSCU	Number	RSCU
Ala	GCA	307	1.13	312	1.12	Leu	CUA	284	0.81	271	0.80
	GCC	141	0.52	153	0.55		CUC	126	0.36	117	0.35
	GCG	127	0.47	131	0.47		CUG	125	0.35	118	0.35
	GCU	514	1.89	518	1.86		CUU	438	1.24	433	1.28
Arg	AGA	363	1.83	342	1.80		UUA	705	2.00	673	1.99
	AGG	115	0.58	112	0.59		UUG	435	1.24	421	1.24
	CGA	274	1.38	261	1.37	Phe	UUC	380	0.67	370	0.70
	CGC	81	0.41	80	0.42		UUU	753	1.33	681	1.30
	CGG	85	0.43	84	0.44	Pro	CCA	239	1.13	229	1.12
	CGU	275	1.38	263	1.38		CCC	152	0.72	151	0.74
Asn	AAC	211	0.42	191	0.43		CCG	126	0.60	111	0.54
	AAU	783	1.58	688	1.57		CCU	326	1.55	324	1.59
Asp	GAC	164	0.39	157	0.40	Ser	AGC	94	0.39	90	0.38
	GAU	668	1.61	624	1.60		AGU	305	1.26	300	1.28
Cys	UGC	56	0.50	56	0.50		UCA	293	1.21	272	1.16
	UGU	170	1.50	166	1.50		UCC	209	0.86	210	0.89
Gln	CAA	573	1.59	528	1.56		UCG	130	0.54	124	0.53
	CAG	150	0.41	147	0.44		UCU	424	1.75	412	1.76
Glu	GAA	838	1.54	753	1.53	TER	UAA	32	1.78	31	1.75
	GAG	248	0.46	232	0.47		UAG	12	0.67	14	0.79
Gly	GGA	548	1.58	551	1.58		UGA	10	0.56	8	0.45
	GGC	146	0.42	143	0.41	Thr	ACA	311	1.26	284	1.20
	GGG	210	0.61	210	0.60		ACC	163	0.66	163	0.69
	GGU	480	1.39	487	1.40		ACG	97	0.39	97	0.41
His	CAC	107	0.43	100	0.42		ACU	416	1.69	400	1.69
	CAU	393	1.57	375	1.58	Trp	UGG	354	1.00	340	1.00
Ile	AUA	607	1.00	547	0.97	Tyr	UAC	155	0.40	151	0.42
	AUC	310	0.51	296	0.52		UAU	614	1.60	576	1.58
	AUU	910	1.49	854	1.51	Val	GUA	417	1.56	420	1.58
Lys	AAA	850	1.54	706	1.54		GUC	117	0.44	111	0.42
	AAG	257	0.46	210	0.46		GUG	137	0.51	137	0.52
Met	AUG	454	1.00	442	1.00		GUU	399	1.49	393	1.48

1.3 Neutrality plot analysis

To evaluate the effect of mutation pressure and natural selection on codon bias in the chloroplast genome of *Platycarya*, the correlation between the composition of the first, second and third bases of codon was analyzed, and a neutral map was drawn. Among the 53 chloroplast coding genes, only 11 genes of *P. longips* distributed along the diagonal line or fell on the diagonal line, the regression coefficient was 0.2004, the contribution rate was 20.04%, but the contribution rate of natural selection pressure was 79.96%; only 13 genes of *P. strobilacea* distributed along the diagonal or on the diagonal, the regression coefficient was 0.2906, the contribution rate of mutation pressure was 29.06%, and the contribution rate of selection pressure was 70.94%. The correlation between GC_{12} and GC_3 was not strong, indicating that mutation pressure had little influence on codon bias, the codon of chloroplast genome was affected by both natural selection and mutation pressure, but natural selection played a greater role (Figure 1).

1.4 ENC-plot analysis

ENC value was calculated to evaluate the codon preference of the protein-coding sequence of the chloroplast genome of *Platycarya*. The standard curve in the ENC-plot analysis indicated that codon preference was completely determined by gene mutation. The encoding genes of *Platycarya* chloroplast genome were plotted as shown in Figure 2, only a few genes (such as matK, rps4, ndhI, ndhC and rpl14) were found in the two species were distributed along or near the standard curve, which indicated that the codon preference of these genes was mainly affected by mutation. However, most of gene distribution in the lower part of standard curve, indicating that the ENC values of most genes were significantly different from the expected values, the codon preference of these genes was affected by other factors, such as natural selection and genes Length.

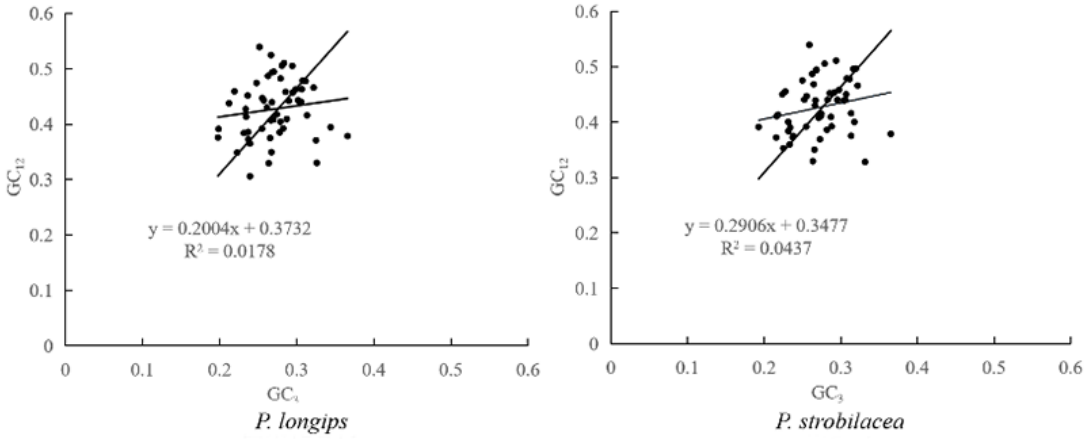

Figure 1 Neutrality plot analysis of *Platycarya* cp genomes

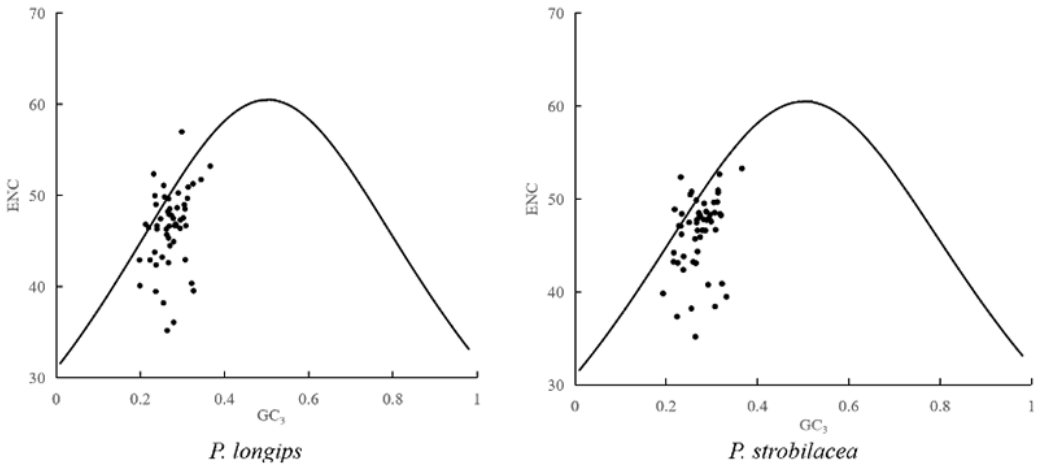

Figure 2 ENC-plot analysis of *Platycarya* cp genomes

1.5 PR2-plot analysis

By drawing the PR2-plot diagram to further analyze the factors affecting the codon preference of the chloroplast genome of the genus *Platycarya*, the coding genes of the chloroplast genome are not evenly distributed in the four areas, and most of the genes are located at the bottom right of the plan, indicating that the use of the third base of the codon in the chloroplast genome is unbalanced, the composition of the third base is T>A, G>C. If the codon preference is only affected by gene mutations, the frequency of four bases will be equal in PR2 plot plot analysis, indicating that most of the genes in the chloroplast genome of *Platycarya* are affected by natural selection or other factors. PR2-plot analysis can only reflect the factors that affect codon usage patterns, Therefore, it is necessary to further study the influence of mutation pressure and natural selection on codon preference (Figure 3).

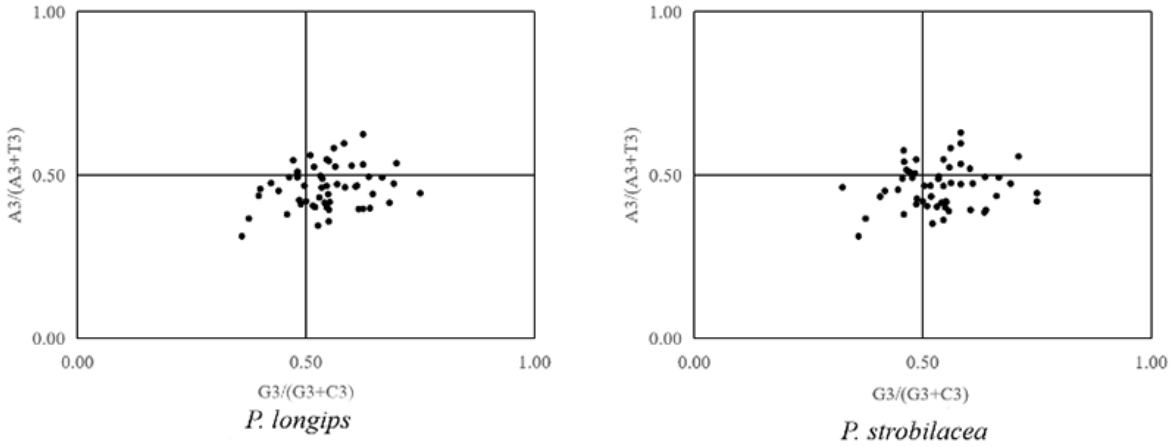

Figure 3 PR2 bias plot analysis of *Platycarya* cp genomes

1.6 Determine the optimal codon

The value of ENC is an important index to reflect the degree of codon preference, usually related to the amount of gene expression. When the ENC value is small, the codon preference of high expression gene is strong, and when the ENC value is large, the codon preference of low expression gene is weak. Based on the ENC value, five high expression genes ycf2, rp12, rps19, ndhB, atpF and five low expression genes cemA, rps18, petD, rps14 and rps15 were selected to establish gene library. Then the RSCU values of the synonymous codons of the two groups were calculated respectively, with RSCU > 1 was the high-frequency codon, ΔRSCU ≥ 0.08 was identified as a superior high expression of chloroplast genome codes, will meet at the same time RSCU > 1 and ΔRSCU ≥ 0.08 as the optimal codons, finally 18 optimal codon were selected, of which seven at the end of A, nine with U, only one at the end of C and G respectively, indicating that the chloroplast genome codon prefer to end with A/U in *Platycarya* (Table 5).

2 Discussion

Chloroplast is a kind of special organelle for photosynthesis in plant cells, which has a complete gene expression system independent of nuclear genome, therefore, chloroplast genome plays an important role in plant phylogeny and taxonomic research. The GC content of the genomes in most species is conserved, and the GC content of the nuclear genome protein-coding genes usually higher than 40%, but the GC content in the chloroplast genome usually less than 40% (Liu et al., 2020). In this study, 53 protein-coding sequences of the chloroplast genomes in *Platycarya* were studied, and the characteristics of the codon usage of the chloroplast genomes were analyzed. The analysis results showed that the codons of the chloroplast genomes of *Platycarya* mainly end with A/U. And there is no significant correlation between the composition of the first、second and third base; the use of the third base of the codon has great A/U preference, which is similar to the results of codons study on chloroplast genome of *persimmon* (Fu et al., 2017), *camphor* (Qin et al., 2016) and *Masson pine* (Ye et al., 2018), but opposite to the result of *Eucommia* transcriptome analysis (Liu et al., 2016), and in the study of *tobacco* and *Arabidopsis thaliana*, the content of GC3 was higher than or similar to that of GC2 (Li et al., 2016).

Table 5 Optimal codons in chloroplast genome of *Platycarya*

Amino acid	Codon	High expression gene		Low expression gene		ΔRSCU
		Number	RSCU	Number	RSCU	
Phe	UUU*	24	1.30	97	1.02	0.28
	UUC	13	0.70	93	0.98	-0.28
Tyr	UAU*	17	1.62	77	1.52	0.09
	UAC	4	0.38	24	0.48	-0.09
His	CAU	10	1.43	65	1.63	-0.20
	CAC*	4	0.57	15	0.38	0.20
Gln	CAA***	18	2.00	79	1.40	0.60
	CAG	0	0.00	34	0.60	-0.60
Asn	AAU	24	1.50	142	1.53	-0.03
	AAC	8	0.50	44	0.47	0.03
Lys	AAA*	38	1.55	118	1.28	0.27
	AAG	11	0.45	66	0.72	-0.27
Asp	GAU	24	1.60	125	1.63	-0.03
	GAC	6	0.40	28	0.37	0.03
Glu	GAA**	32	1.78	109	1.28	0.50
	GAG	4	0.22	61	0.72	-0.50
Cys	UGU***	7	2.00	25	1.32	0.68
	UGC	0	0.00	13	0.68	-0.68
Val	GUU*	12	1.50	36	1.33	0.17
	GUC	2	0.25	17	0.63	-0.38
	GUA***	14	1.75	31	1.15	0.60
	GUG	4	0.50	24	0.89	-0.39
Pro	CCU	8	0.94	37	1.38	-0.44
	CCC***	15	1.76	21	0.79	0.98
	CCA	6	0.71	32	1.20	-0.49
	CCG	5	0.59	17	0.64	-0.05
Thr	ACU***	14	1.87	44	1.35	0.51
	ACC	4	0.53	26	0.80	-0.27
	ACA	10	1.33	41	1.26	0.07
	ACG	2	0.27	19	0.58	-0.32
Gly	GGU***	16	1.78	30	0.89	0.89
	GGC	2	0.22	13	0.39	-0.16
	GGA	14	1.56	64	1.90	-0.34
	GGG	4	0.44	28	0.83	-0.39
Ala	GCU	9	1.38	34	1.72	-0.34
	GCC	1	0.15	13	0.66	-0.50
	GCA***	13	2.00	17	0.86	1.14
	GCG	3	0.46	15	0.76	-0.30
Ile	AUU**	35	1.75	104	1.28	0.47
	AUC	9	0.45	55	0.68	-0.23
	AUA	16	0.80	84	1.04	-0.24
Leu	UUA***	30	1.96	65	1.24	0.72
	UUG**	28	1.83	73	1.39	0.44
	CUU	16	1.04	72	1.37	-0.33
	CUC	4	0.26	27	0.51	-0.25
	CUA	12	0.78	52	0.99	-0.21
	CUG	2	0.13	26	0.50	-0.36
Ser	UCU	11	1.29	75	1.60	-0.31
	UCC	10	1.18	55	1.17	0.01
	UCA*	12	1.41	59	1.26	0.15
	UCG	5	0.59	33	0.70	-0.12
	AGU***	12	1.41	42	0.90	0.51
	AGC	1	0.12	17	0.36	-0.25
Arg	CGU***	17	2.32	26	0.83	1.49
	CGC	0	0.00	13	0.41	-0.41
	CGA	8	1.09	45	1.44	-0.35
	CGG	3	0.41	21	0.67	-0.26
	AGA	12	1.64	56	1.79	-0.15
	AGG	4	0.55	27	0.86	-0.32

Note: *indicates ΔRSCU≥≥0.08; **indicates ΔRSCU≥0.30; ***indicates ΔRSCU≥0.50; __indicates RSCU>1

In general, the GC content of position in codon 1, 2, 3 of chloroplast genome is different, and GC1 > GC2 > GC3. The base composition of codons in dicot plants prefers to end with A/T, while that of monocotyledons shows extreme G/C preference (Wang and Roossinck, 2006), this is consistent with the study on the codon base composition of the chloroplast genome in this study.

The number of effective codons ranges from 20 to 61. When the ENC value is low, it means that the codon has a higher preference, the ENC value equal to 35 is usually considered as a critical value for judging the strength of the codon preference. In the research of *Platycarya*, it is found that the ENC value of the chloroplast genome ranges from 35.19 to 56.98, and the ENC values of all genes are greater than 35, indicating that the codons of the chloroplast *Platycarya* genome are weak in Codon usage. There are a total of 30 codons with RSCU>1 in *Platycarya* chloroplast genome, 29 of which end in A/U, and 18 optimal codons are selected by using ENC value and ΔRSCU as indexes. The results of the analysis showed that the codon preference of the chloroplast genome of *Platycarya* ends in A/U, which is the same as the result of the base composition analysis, is similar to previous studies on *Diospyros* plants (Fu et al., 2017) and *Populus alba* (Zhou et al., 2008).

Studies have shown that the factors influencing codon usage preference include mutation bias, natural selection (Marais et al., 2003), GC content (Sun et al., 2009; Hunt et al., 2014), synonymous substitution rate, tRNA abundance, gene length and expression level (Pop et al., 2014), among which gene mutation and natural selection are the most important factors (Morton, 2003; Prabha et al., 2017). The analysis of neutral plots, ENC-plot, and PR2-plot shows that the codon preferences of the chloroplast genomes of the Platycarya were affected by mutations and natural selection at the same time, but the influence of natural selection is greater, and it is inferred that Similar evolutionary patterns may exist between two closely related species. This conclusion is consistent with the conclusions of previous studies (Xu et al., 2017), and it may be related to the relative conservation of chloroplast genome evolution. Through the field investigation, we found that the ecological environment of two *Platycarya* is very similar, only distributed in limestone habitat, and both species are the dominant species in Karst forest ecosystem, the same ecological environment may be the reason for similar evolution direction of the genus. At present, there are still some dispute on the taxonomic definition of the genus. Some people think that there is no significant difference between the two species in morphology and statistics, and there are obvious transitional characters between them, *Platycarya* is determined as a single genus (Zhang Li et al., 2011, National Symposium on Systematics and Evolutionary Botany and youth, 2011). However, it is not entirely accurate to define the taxonomic status of plants only from morphology, in this study, we analyzed the base composition of chloroplast genome, the degree of codon preference, and the influencing factors of codon preference, although there are similarities between the two species, they are not identical, so it is not supported to classify the two species into the same species.

This study is the first time to comprehensively and systematically introduce the factors affecting the synonymous codons of the chloroplast genome of *Platycarya*, and also systematically discuss the codons use patterns. The results of this study can provide theoretical basis for modification of foreign codons, prediction of chloroplast genes and determination of unknown genes.

3 Materials and Methods

3.1 Materials

The leaves of *P. longipes* were collected from Fengxiang mountain (106 ° 38 ′ 0.59 ″ E, 26 ° 22 ′ 54.45 ″ N, altitude at 1205 m) of Huaxi, Guiyang. The collected fresh leaves were stored in dry ice and sent for chloroplast genome sequencing. After sequencing, the chloroplast genome data were submitted to NCBI database, and the accession number was MT032191. The *P. strobilacea* chloroplast genome was obtained from NCBI database with accession number kx868670. In order to reduce the analysis error, the coding sequence of chloroplast genome was screened, and the length of CDs less than 300 bp and repeated CDs sequences were removed, finally, 53 CDs sequences were obtained for the two species for analysis.

3.2 Analysis of codon composition

Based on 53 CDs sequences obtained from two species after screening, the effective number of codon (ENC) and relative synonymous codon usage (RSCU) were analyzed by software codonw1.4.2, online software cusp (http://emboss.toulouse.inra.fr/cgi-bin/emboss/cusp), the total GC content of codon and the GC content of position 1, 2 and 3 were calculated using Excel and SPSSAU (https://spssau.com/front/spssau/index.html).

3.3 Neutral plot analysis

Neutral plot is a method to study the influence of mutation pressure and natural selection on codon usage patterns (Wang et al., 2018). Synonymous codon mutations usually occur at the third base of codon, and the first and second bases usually have nonsynonymous codon mutations, and the mutation probability is low. Taking the average of GC_1 and GC_2 as the ordinate and GC_3 as the abscissa, a neutral map is drawn. Each point in the graph represents a gene, if all the points in the graph are distributed along a diagonal line, the base composition of the three codons is not different, and the codon usage is only affected by the mutation pressure, if the correlation between GC_{12} and GC_3 is very low, it indicates that natural selection is the main determinants of codon usage patterns.

3.4 ENC-plot analysis

Effective number of codon (ENC) is used to describe the degree to which codon usage deviates from random selection. The ENC value ranges from 20 and 61. When the ENC value is 20, it means that only one codon is used for each amino acid, showing extreme preference; when the ENC value is 61, means that the use of codon is random and there is no preference (Wang et al., 2018). Generally, ENC = 45 is used as the criterion to distinguish the degree of bias (Ye, 2018). Taking ENC value as ordinate and GC_{3s} as abscissa, a two-dimensional scatter plot is drawn. The expected value of codon preference completely determined by mutation is taken as the standard curve, the calculation formula of standard curve is: $ENC = 2 + GC_{3S} + 29 / [GC_{3s}^2 + (1\text{-}GC_{3s})2]$. When the gene is located or close to the standard curve, the codon preference is mainly affected by mutation, when the distance from the standard curve is far from the standard curve, natural selection has a greater impact on the codon usage pattern.

3.5 PR2-plot analysis

PR2-plot is used to analyze the composition of the third base of codon, A3 / (A3 + T3) used as ordinate and G3 / (G3 + C3) as abscissa to draw scatter plot, the central position of scatter plot (when A = T, C = G) represents the usage of codon without preference.

3.6 Optimal codon determination

In order to determine the optimal codon, five samples with the highest and lowest ENC values were taken as the low and high expression groups, respectively. The RSCU difference between the low expression group and the high expression group (Δ RSCU $\geq$ 0.08) was defined as the high expression superior codon, and the RSCU value greater than 1 was defined as the high-frequency codon, the optimal codon is determined when the two conditions are satisfied.

Authors' contributions

Wang Xiaoshuang is responsible for writing the first draft of the article and analyzing the experimental data. Wang Yanqing and Li Shihong are responsible for data collection and collation. Liu Yingliang is responsible for the conception and experimental design of the paper. Zhu bin is responsible for the guidance and revision of the experiment. All authors read and approved the final manuscript.

Acknowledgements

Supported by the National Natural Science Foundation of China (No. 31760124); The Joint Fund of the National Natural Science Foundation of China and the Karst Science Research Center of Guizhou Province (Grant No. U1812401).

References

Freitas A., Anunciação R., D'Oliveira Matielo C.B., and Stefenon V.M., 2018, Chloroplast DNA: A promising source of information for plant phylogeny and traceability, Journal of Molecular Biology and Methods, 1(1): 1-4

Fu J.M., Suo Y.J., Liu H.M., and Tan X.F., 2017, Analysis on codon usage in the chloroplast protein-coding genes of *Diospyros* spp, Jingjilin Yanjiu (Nonwood Forest Research), 35(2): 38-44

Hunt R.C., Simhardri V.L., Iandoli M., Sauna Z.E., and Sarfaty K., 2014, Exposing synonymous mutations, Trends Genet., 30(7): 308-321
https://doi.org/10.1016/j.tig.2014.04.006

Kong W., and Yang J., 2016, The complete chloroplast genome sequence of morus mongolicaand a comparative analysis within the Fabidae clade, Curr Genet., 62(1): 165-172
https://doi.org/10.1007/s00294-015-0507-9

Li G., Pan Z., and Gao S., 2019, Analysis of synonymous codon usage of chloroplast genome in *Porphyra umbilicalis*, Genes & Genomics, 41(10): 1173-1181
https://doi.org/10.1007/s13258-019-00847-1

Li N., Li Y., and Zheng C., 2016, Genome-wide comparative analysis of the codon usage patterns in plants, *Genes & Genomics*, 38(8): 723-731
https://doi.org/10.1007/s13258-016-0417-3

Liu H.B., Lu Y.Z., Lan B.L., and Xu J.C., 2020, Codon usage by chloroplast gene is bias in *Hemiptelea davidii*, Journal of Genetics, 99(1): 8
https://doi.org/10.1007/s12041-019-1167-1

Liu H.M., Wuyun T.N., and Du H.Y., 2016, Analysis of characteristic of codon usage of *Eucommia ulmoides* Transcriptome, Beifang Yuanyi (Northern Horticulture), 40(13):85-89

Liu JK., Ying M., and Wang Q., 2016, Content determination of gallic acid in infructescence of *Platycarya Strobilacea*, Zhongguo Yaoye (China Pharmaceuticals) 27(6): 4-7

Long S.Y., Yao H.P., Wu Q., and Li G.L., 2018, Analysis of compositional bias and codon usage pattern of the coding sequence in *Banna virus* genome, Virus Res., 258: 68-72
https://doi.org/10.1016/j.virusres.2018.10.006

Marais G., Mouchiroud D., and Duret L., 2003, Neutral effect of recombination on base composition in Drosophila, Genetics Res., 81(2): 79-87
https://doi.org/10.1017/S0016672302006079

Morton B.R., 2003, The role of context-dependent mutations in generating compositional and codon usage bias in grass chloroplast DNA, J. Mol. Evol., 56(5): 616-629
https://doi.org/10.1007/s00239-002-2430-1

Niu Y., Xu Q., Wang Y.D., Dai L.L., Zhuang J., and Zhao Y.L., 2018, An analysis of codon use bias of chloroplast genome of *Rosa odorata* var.*gigantea*, Xibei Linxueyuan Xuebao (Journal of Northwest Forestry University), 33(3): 123-130

Pop C., Rouskin S., Ingolia N.T., Han L., and Phizicky E.M., 2014, Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation, Mol. Syst. Biol., 10(12): 770
https://doi.org/10.15252/msb.20145524

Prabha R., Singh D.P., Sinha S., Ahmad K., and Rai A., 2017, Genome-wide comparative analysis of codon usage bias and codon context patterns among cyanobacterial genomes, Mar. Genomics, 32: 31-39
https://doi.org/10.1016/j.margen.2016.10.001

Qi Y.Y., Xu W.J., and Xing T., 2015, Synonymous codon usage bias in the plastid genome is unrelated to Gene structure and shows evolutionary heterogeneity, Evol. Bioinform., 11(11): 65-77
https://doi.org/10.4137/EBO.S22566

Qin Z., Cai Z.Q., Xia G.M., and Wang M.C., 2013, Synonymous codon usage bias is correlative to intron number and shows disequilibrium among exons in plants, BMC Genomics, 14(1): 56-56
https://doi.org/10.1186/1471-2164-14-56

Qin Z., Zheng Y.J., Gui L.J., Xie G.A., and Wu Y.F., 2018, Codon usage bias analysis of chloroplast genome of camphora tree (*Cinnamomum camphora*), Guangxi Zhiwu (Guihaia), 38(10): 90-99

Suiura M., 1992, The chloroplast genome, *Plant Molecular Biology*, 19(1): 149-168
https://doi.org/10.1007/BF00015612

Sun Z., Ma L., Murphy R., Zhang X.S., and Huang D.W., 2009, Analysis of codon usage on *Wolbachia pipientis w*Mel genome, Zhongguo Kexue (Science in China), 39(10): 948-953

Wang L., Roossinck M J., 2006, Comparative analysis of expressed sequences reveals a conserved pattern of optimal codon usage in plants, Plant Molecular Biology, 61(4): 699-710
https://doi.org/10.1007/s11103-006-0041-8

Wang L., Pan W.D., and Zhou S., 2012, Structural mutations and reorganizations in chloroplast genomes of flowering plants, Xibei Zhiwu Xuebao (Acta Botanica Boreali-Occidentalia Sinica), 32(6): 1282-1288

Wang M.Y., Liu J.T., and Hou N., 2010, Determination of gallic acid in the fruit sequence of *Platycarya strobilacea* by RP-HPLC, Zhongguo Yaoshi (Chinese Pharmacist), 13(3): 68-69

Wang P.L., Wu S.C., Yang L.P., Wang H.Y., Chen N.M., and Zhang Z.Y., 2019, Analysis of codon preference of eucalyptus grandis chloroplast genome, Guangxi Zhiwu (Guihaia), 39(12): 1583-1592

Wang H.G., Tao M., and Wen W.Q., 2018, Analysis of synonymous codon usage bias in helicase gene from *Autographa californica* multiple nucleopolyhedrovirus, Genes & Genomics, 40(7): 767-780
https://doi.org/10.1007/s13258-018-0689-x

Wu X.M., Wu S.F., Ren D.M., Zhu Y.P., and He F.C., 2007, The analysis method and progress in the study of codon bias, Yichuan (Hereditas), 29(4): 420
https://doi.org/10.1360/yc-007-0420

Xu C., Dong W.P., Li W.Q., Lu Y.Z., Xie X.M., Jin X.B., Shi J.P., He K.H., and Suo Z.L., 2017, Comparative analysis of six *Lagerstroemia* complete chloroplast genomes, Front. Plant Sci., 8(15): 15-27
https://doi.org/10.3389/fpls.2017.00015

Xu Y., Luo XJ., and Liu YL., 2019, Analysis of the age structure and spatial distribution pattern of the population in different regions of Guizhou, Fenzi Zhiwu Yuzhong (Molecular Plant Breeding), 17(11): 3769-3777

Yang Y., 2010, Determination of ascorbic acid content in *Platycarya longipes* wu by spectrophotometery, Anhui Nongye Kexue (Journal of Anhui Agricultural Sciences), 38(18): 9523-9526

Ye Y.J., Ni Z.X., Bai T.D., and Xu L.A., 2018, The analysis of chloroplast genome codon usage bais in *Pinus massoniana*, Jiyinzuxue Yu Yingyong Shengwuxue (Genomics and Applied Biology), 37(10): 4464-4471

Zhang J., Jiang Z., and Su H., 2019, The complete chloroplast genome sequence of the endangered species *Syringa pinnatifolia* (Oleaceae), Nord. J. Bot., 37(5): 2201-2212
https://doi.org/10.1111/njb.02201

Zhao Y.C., Zheng H., Xu A.X., Yan D.H., Jiang Z.J., Qi Q., and Sun J.C., 2016, Analysis of codon usage bias of envelope glycoprotein genes in nuclear polyhedrosis virus (NPV) and its relation to evolution, BMC Genomics, 17(1): 677-677
https://doi.org/10.1186/s12864-016-3021-7

Zhou M., Long W., and Li X., 2008, Analysis of synonymous codon usage in chloroplast genome of *Populus alba*, Journal of Forestry Research, 19(4): 293-297
https://doi.org/10.1007/s11676-008-0052-1