Genome-Wide Analysis of Synonymous Codon Usage Bias in *Cycas panzhihuaensis*

Yongjian Luo[1,2,3]& Ru Wang[1,3], Qing Li[2],Jun Liu[2]*, Songquan Song[2], Zhijun Deng[1,3]*

1 Hubei Key Laboratory of Biologic Resources Protection and Utilization (Hubei Minzu University), Enshi,The Plant Germplasm Resources Laboratory, School of Forestry and Horticulture, Hubei Minzu University, Enshi, P. R. China

2 Guangdong Key Laboratory for Crop Germplasm Resources Preservation and Utilization, Agro-biological Gene Research Center, Guangdong Academy of Agricultural Sciences, Guangzhou, Guangdong,China

3 Research Center for Germplasm Engineering of Characteristic Plant Resources in Enshi Prefecture (Hubei Minzu University), Enshi, P. R. China

* Correspondence:dengzhijun@hbmzu.edu.cn; Jun Liu liujun@gdaas.cn

Abstract: Codon use bias is an important characteristic in the process of genetic information transmission of species. With the publication of the genome of *Cycas*, it is of great significance to investigate the codon use bias for understanding the genetic evolution of this species and for molecular breeding. In this study, Perl language and CodonW 1.4.2 software was used to systematically analyze the coding gene sequences of the *Cycas* genome, obtain the characteristics of codon use, and identify the sources of variation affecting codon use. The results showed that GC3s content, GCall content, and ENC value were 46.75%, 47.67%, and 52.8 respectively. As the first seed plant, the GC3 content of *Cycas* is similar to that of most gymnosperms and most dicotyledons, but considerably different from that of monocotyledons.RSCU value was greater than 1, among which 21 codons ended in U or A. Enc-plot analysis, PR2-plot neutral plot analysis, and correlation coefficient analysis of Axis1 with GC3s and CAI revealed that the codon bias was mainly affected by natural selection, but also by mutation pressure and other factors. Thirty-one optimal codons of *Cycas* genes were screened. This research can provide a reference for the phylogeny and genome codon evolution of *Cycas*.

Keywords: RSCU;ENc-GC3s plot; synonymous codon usage bias; *Cycas*;CAI

1 Introduction

Genetic code is the link between nucleic acid and protein and plays an important role in the transmission of genetic information in organisms. A total of 64 codons encode 20 amino acids and 3 termination signals, respectively[1]. Except for Met and Trp, which are encoded by only one codon, other amino acids all have multiple synonymous codons . In the process of protein translation, the usage probability of synonymous codons is different. A species of a gene tends to use one or more specific synonymous codons, which is called synonymous codon usage bias (SCUB) [2]. SCUB originates from mutation, natural selection, and genetic drift, and is affected by genome composition, GC content, gene length and expression level, location and background of codons in genes, gene recombination rate, mRNA folding, tRNA abundance, etc. SCUB is prevalent in prokaryotes and eukaryotes. Gene expression and cell function are determined through biochemical processes such as RNA processing, protein translation, and folding[2–4]. SCUB research, therefore, not only can reflect the origin of the species or gene, but the mutation model and evolution can also reveal the phylogenetic relationship between biological, horizontal gene transfer, gene molecular evolution,

 doi:10.20944/preprints202205.0208.v1

and identification of driving evolutionary selection pressure, and can promote gene in the transformed plants by codon optimization of expression, and then promote the development of genetically modified (gm) crops[5].

Cycas are one of the oldest species of seed plants on earth[6]. They first appeared in the Permian period of the earth about 280 million years ago[7,8]. The mass extinction of *Cycas* in the northern hemisphere occurred during the quaternary glaciation[7]. Due to the Qinghai-Tibet Plateau, the Qinling mountains and other barriers in China, some *Cycas* distributed in South China were spared[9]. The existing *Cycas* evolved from the few descendants left by many ancestral groups, so the *Cycas* are known as "Livingfossil"[10]. The roots, leaves, seeds and trunks of *Cycas* have different medicinal effects, and leaf extracts have anti-cancer activities[11–13]. Therefore, *Cycas* has important development and utilization value. The study of *Cycas* is of great scientific value in the exploration of paleontology, palaeoecology, palaeogeology, palaeogeography and the origin and evolution of seed plants[14].

Recently, scientists published a complete genome map of *Cycas panzhihuaensis* [15]. This means that the "last piece of the puzzle" of the genome evolution of seed plants has been successfully completed, while there is no related research on the SCUB genome of *C. panzhihuaensis*. In this research, bioinformatics methods were used to study the SCUB genome of *C. panzhihuaensis*, providing reference for phylogeny, genetic evolution, molecular breeding, and species conservation of *Cycas*.

2. Methods

2.1 Database preparation

Publicly available *C. panzhihuaensis* genome sequences that contain complete coding sequences (CDSs) were obtained from the CNGB database (https://ftp.cngb.org/pub/CNSA/data1/CNP0001756/CNS0381401/CNA0022716/)In total, 32,353 coding sequences (CDS) were obtained. To improve the quality of sequencing and to minimize sampling errors, the sequences containing correct initiation and termination codons remained and sequences containing internal termination codons were eliminated by using Select CDS script[16]. In addition, only the sequences longer than 100 amino acids in length were held for further analysis. The final sequence collection containing 30,007 CDS was used for further codon usage analyses.

2.2. Indices of codon usage and codon bias

Using Mobyle software (http://www.molbiol.ox.ac. UK/cu, version 1.4.2) calculate each gene sequence of nucleobases[17], and statistical indicators: content of each nucleobases of the third codon (A3s, U3s, C3s, G3s)； the GC content at the first codon position of synonymous codons （GC1）,the GC content at the second codon position of synonymous codons （GC2）； Overall GC content of codon （GCall）； the average GC content at the first and second codon position of synonymous codons (GC12)； the GC content at the third codon position of synonymous codons (GC3s),Overall GC content of codon （GCall）.In addition, codon usage indexes of all genes were calculated using CodonW program[18].For example, effective number of codon(ENC), codon adaptation index(CAI), relative synonymous codon usage (RSCU). The ENC value is used to evaluate the extent to which codon use deviates from random selection[19], which describes the extent to which the use of synonymous codons is unbalanced in the genes or genomes of a particular

 doi:10.20944/preprints202205.0208.v1

species. The ENC value ranges from 20 to 61. A smaller ENC value indicates a stronger SCUB [20]. The CAI value evaluates the degree to which the synonymous codon applied deviates from the favored synonymous codon of the highly expressed gene. The CAI value has a certain correlation with the gene expression level, and its value is between 0 and 1. The higher the CAI value is, the stronger the codon usage bias is and the higher the gene expression level[21].The RSCU value refers to the ratio between the usage frequency of a particular codon and the anticipated frequency of unbiased use and is also an effective indicator to measure the degree of codon bias. RSCU=1 implies that there is no bias in the use of the codon. RSCU> 1 indicates that the frequency of the codon is too high. RSCU < 1 indicates that the frequency of the codon is low[21,22].

2.3. ENc-GC3s plot and PR2-Bias plot analysis

In order to explore the relationship between SCUB and nucleobase composition, an ENC-plot(ENC vs GC3s) plotting analysis was conducted [22].A scatter plot was drawn with GC3 as abscissa and ENC as ordinate.The standard ENC value of each gene was calculated by ENC standard curve formula "ENC = 2 + GC3 + 29/(GC32 + (1-GC3)^2)", and then the standard curve was drawn in the scatter plot with GC3s as abscissa and standard ENC as ordinate. If the location of each gene in the figure is distributed near the standard curve, it indicates that mutation is the main influencing factor of SCUB.If the location distribution of genes is far from the standard curve, it indicates that SCUB is mainly affected by natural selection and other factors (丁锐等, 2021).

PR2-plot bias analysis is used to explore the influence of mutation and natural selection on SCUB. The center point of PR2-plot indicates that there is no mutation or selection effect bias between two complementary chainsthe of A gene, that is, A = T and G = C. The vector distribution from the central point to other loci reflects the direction and level of bias of the gene. It is generally believed that the ratio of A/T and C/G in the degenerate codon of A gene or genome is balanced under the pressure of A single mutation **[24]**. By referring to the method of Duan (2021), the contents of, A, T, C, and G at the third nucleobases of the codon were calculated, respectively, and the Parity Rule 2 bias (PR2-Bias) plot analysis was conducted using the value of G3/(G3 + C3) as the horizontal coordinate and the value of A3/(A3 + T3) as the vertical coordinate **[22]**.

2.4. Neutrality plot analysis

Neutral plots (GC12vs.GC3) were used to study the influence of mutation pressure and natural selection on codon usage patterns (Sueoka., 2021). A scatter plot was drawn with GC12 as ordinate and GC3 as abscissa, in which each point represented the location of a gene. The correlation between GC12 and GC3 in the Davidia involucrate genome was calculated using Perl scripting language. The slope of the curve regression is close to zero and is strongly influenced by natural selection. A slope close to 1 indicates that the codon usage bias is completely influenced by the directional mutation pressure representing complete neutralit[25].

2.5. Correspondence analysis of codon usage

Correspondence analysis (COA) has been widely used to explore the variation in synonymous codon usage among genes.COA-plot is a sophisticated multivariate statistical technique in which codon usage data (59 codons excluding Met, Trp, and stop codons) was plotted in a multidimensional space of 59 axes.The plot was then used to identify the axes that represent the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 May 2022 doi:10.20944/preprints202205.0208.v1

most prominent factors contributing to variation among genes.Based on the results of codon usage variation, correlation analysis between axis 1 and CAI,ENC,GCall;GC12,GC3s, were carried out by R (Version 4.12)[26].

2.6. Determination of optimal codons

The selected CDS sequences were sorted from high to low according to CAI value, and 5% of genes at both ends were selected as high and low expression libraries. Codon usage was compared using chi-squarered contingency test of the 2 groups, and codons whose frequency of usagwasre significantly higher ($P < 0.01$)isn highly expressed genes than in genes with low level of expression would be defined as the optimal codons[21].

3 Results

3.1. Nucleotide composition of *C. panzhihuaensis* and codon bias

SCUB for a single type of codon is greatly influenced by the overall nucleotide content of the genome[27,28]. Among the 30007 CDS in *C. panzhihuaensis panzhihuaensis*, the nucleotide content of A varies from 4.32% to 69.72%, with a mean value of 30.43%, the frequency of T is 1.47%~60.49%, with an average value of 36.37%, the proportion of G is 1.73%~ 82.11%, with a mean value of 30.79%, the ratio of *C. panzhihuaensis panzhihuaensis* is 0.82%~69.39%, with a mean value of 26.07%. To further understand the impact of nucleotide contents of *C. panzhihuaensis panzhihuaensis* genes on codon bias, we also calculated the GC contents and GC3s. The results showed that GC content of all genes mainly concentrated in 40% ~ 55% frequency(Figure 1), the nucleotide content of GC varies from 29.51% to 83.45%, with a mean value of 47.67%.This indicated that the gene AT content was higher than GC content in C. palmata, which was similar to that in most plants.The mean value of GC1,GC2 and GC3s was 52.72%, 43.20 and 46.75, respectively, which testified that the GC content at the three codon sites is not uniform and tends to end with A and T nucleobases.

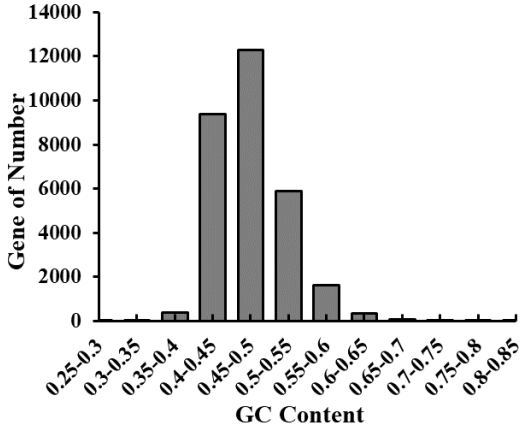

Figure 1 The distribution of GC contents in the CDS of *C. panzhihuaensis*.

We used the neutral plot (GC12 versus GC3) to analyze the relationship between GC12 and GC3, and to determine the main driving force of SCUB. The values of GC12 ranged from 32.21% to 84.08, with an average of 47.96%. The values of GC3 ranged from 19.08 to 89.36.The regression slope of GC12 and GC3 was 0.147(Figure 2), indicating that the SCUB of *C. panzhihuaensis*'s

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 May 2022 doi:10.20944/preprints202205.0208.v1

genome data was mainly influenced by natural selection rather than mutation pressure.

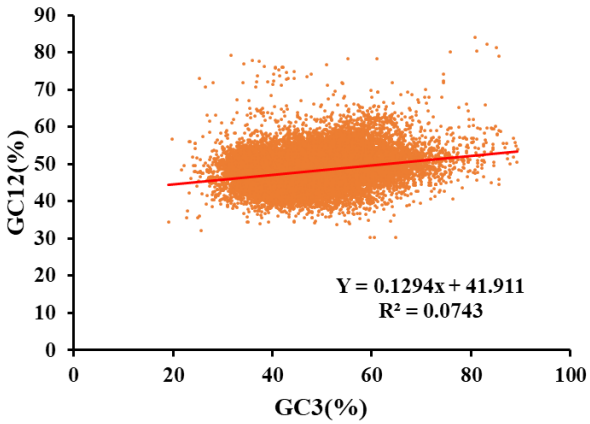

Figure 2. Neutrality plots analysis of GC12 and GC3 of *C. panzhihuaensis*.

3.2. The effect of RSCU and ENC on codon bias

ENC values ranged from 21.43 to 61.00, with an average value of 52.83. ENC results showed that 85.43% of genes had ENC values higher than 50, indicating that the corresponding synonymous codon was used random in encoding various amino acids in genes, and SCUB was low. Meanwhile, RSCU analysis showed (Table 3) that there were 26 codons with RSCU greater than 1, among which 21 ended with A and U, accounting for 80.76%. The results showed that the *C. panzhihuaensis* genome preferred to use synonymous codons ending in A or U, but the codons with RSCU greater than 1 accounted for about half of the 59 codons, which also confirmed the fact that the *C. panzhihuaensis* codon bias was weak.

Table 1. Codon usage in *C. panzhihuaensis*.

AA	Codon	RSCU	AA	Codon	RSCU	AA	Codon	RSCU
Ala	GCA	**1.4452**	Lys	AAA	0.9505	Ser	AGC	0.9114
	GCC	0.8337		AAG	**1.0495**		AGU	0.9488
	GCG	0.4844	Leu	CUA	0.5108		UCA	**1.1663**
	GCU	**1.2367**		CUC	0.7777		UCC	0.8956
Cys	UGC	0.9966		CUG	**1.173**		UCG	0.6317
	UGU	**1.0034**		CUU	**1.2903**		UCU	**1.4461**
Asp	GAC	0.6775		UUA	0.7481	Thr	ACA	**1.3278**
	GAU	**1.3225**		UUG	**1.5001**		ACC	0.8401
Glu	GAA	**1.0593**	Asn	AAC	0.7222		ACG	0.6678
	GAG	0.9407		AAU	**1.2778**		ACU	**1.1642**
Phe	UUC	0.8619	Pro	CCA	**1.2666**	Val	GUA	0.7054
	UUU	**1.1381**		CCC	0.82		GUC	0.6926
Gly	GGA	**1.2987**		CCG	0.5936		GUG	**1.2411**
	GGC	0.8647		CCU	**1.3197**		GUU	**1.3609**
	GGG	0.8394	Arg	AGA	**1.6823**	Tyr	UAC	0.8123
	GGU	0.9972		AGG	**1.4524**		UAU	**1.1877**
His	CAC	0.741		CGA	0.7823	Gln	CAA	0.9715

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 May 2022 doi:10.20944/preprints202205.0208.v1

	CAU	**1.259**	CGC	0.6075	CAG	**1.0285**
Ile	AUA	0.7774	CGG	0.7036		
	AUC	0.8029	CGU	0.7719		
	AUU	**1.4198**				

Notes. The preferentially used codon s are displayed in bold.

3.3. The role of GC3s in the codon bias formation

The curve of the gene above the ENC value is larger, the gene codon use is random, codon bias is comparatively weak, as can be seen from figure 3, *C. panzhihuaensis* in most genetic deviation from the standard curve of ENCexp. This means that in addition to the mutation pressure on codon bias, the influence of the genome is more from the effects of natural selection. Most genes were located below the standard curve, implying that most genes had a strong bias

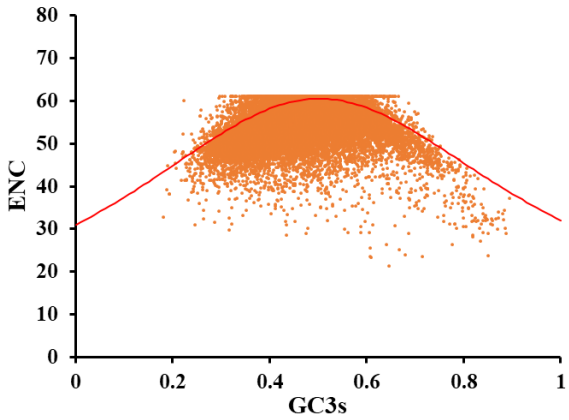

Figure 3. Neutrality plots analysis of GC12 and GC3 of *C. panzhihuaensis*.

Further, we calculated (ENCexp-ENCobs)/ENCexp for all the genes in*C. panzhihuaensis panzhihuaensis* to discriminate the difference between observed and expected ENC values. As shown in Figure. 4 (ENCexp-ENCobs)/ENCexp values for most genes were within 0.05–0.15 which indicated that the most expected ENC values were bigger than actual ENC values.The values of GC3s affected the results of ENC value according to the calculation formula of expected ENC. The distribution frequency of (ENCexp-ENCobs)/ENCexp values was consistent with the results that have been shown in Figure 4, which suggested that the different values of GC3s can affect the result of codon bias.Taken together, the results provide more evidences that GC3s play as a conditional mutational bias.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 May 2022 doi:10.20944/preprints202205.0208.v1

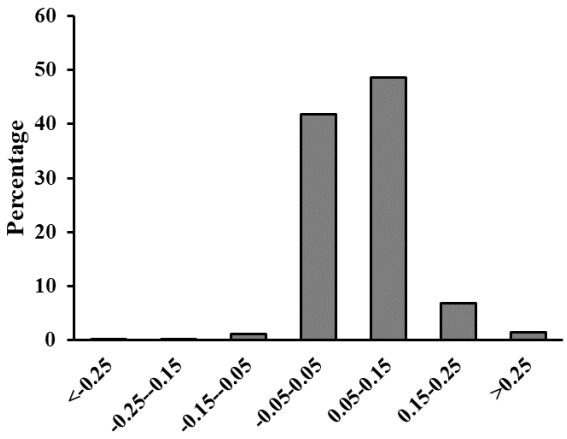

Figure 4. Frequency distribution of the ENC ratio.

3.4. Correspondence analysis

We conducted correspondence analysis (COA) on SCUB values of all genes in C. panzhihuaensis. The results showed that the first four axes accounted for 4.20% of the total contribution rate, and the first, second, third, and fourth axes contributed 18.64%, 11.58%, 10.34%, and 10.64%, respectively, indicating that axis 1 was the main source of variation of codon preference. To verify the effect of GC content on codon bias, a COA-plot (Axis1 vs. Axis2) was constructed with Axis1 as abscissa and Axis2 as ordinate. Green points represent less than 45% GC content, red points more than 45% and less than 60%, and blue points more than 60%. As shown in Figure 5, the blue dots were separated along the primary axis, and the green and red dots were located in the middle of the plot.

Mutation pressure and selection are the main factors affecting codon usage patterns. To analyze the correlation between Axis 1 and GCall, CAI, ENC, GC12, and GC3, and assess the effects of mutation pressure and natural selection on the codon usage patterns of *C. panzhihuaensis*. The results showed that Axis 1 and GC3s (R^2=0.127, P < 0.01), GCall (R^2=0.110, P < 0.01), GC12 (R^2=0.047, P < 0.01), ENC (R^2=0.030, P < 0.01), and CAI(R^2=0.013, P < 0.05) considerably correlated,implying that the codon bias of*C. panzhihuaensis panzhihuaensis* genes were influenced by two main factors including mutational pressure and translational selection.

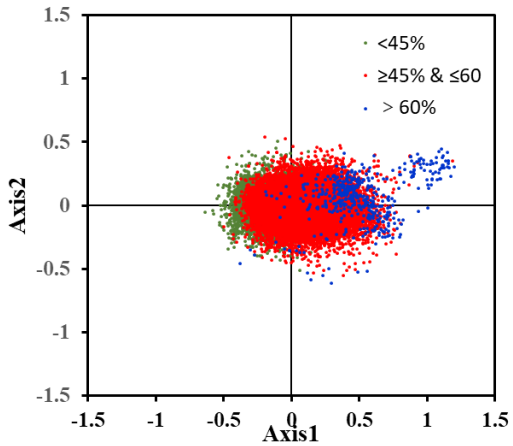

Figure 5. Correspondence analysis of codon usage bias: genes with GC content higher than 60%, within 45%–60% and lower than 45% were plotted as blue, red and green dots, respectively.

Table2 Correlation analysis of*C. panzhihuaensis panzhihuaensis* gene-related parameters.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 May 2022 doi:10.20944/preprints202205.0208.v1

	NEC	GC3s	GC_{all}	GC12	Axis1
CAI	-0.061**	0.159**	0.037**	-0.094**	0.013*
Nc		0.167**	0.111**	0.01	0.030**
GC3s			0.819**	0.292**	0.127**
GC				0.784**	0.110**
GC12					0.047**

Notes. *Significant difference at $p < 0.05$. **Significant difference at $p < 0.001$

3.5. Parity Rule 2 (PR2) plot analysis

The third nucleobases of the codon plays a leading role in codon bias and in the self-protection mechanisms of species evolution[29]. Therefore, whether the third nucleobases AU and GC of the codon are evenly used reflects the influence of mutation pressure on the R.*C. panzhihuaensis* genome. The relationship between the third superpurine (A and G) and pyrimidine (T and C) in the four codon amino acid families was analyzed by PR2 bias plot[30]. Results As shown in Figure 6, A3 and U3(or G3 and C3) frequencies are different in *C. panzhihuaensis*, and the distribution of most points between 0.2 and 0.8 in the figure 6 indicates a low bias in G3/ C3 or A3/ T3. In addition, the plot is divided into four quadrants, 0.5 centered on two axes, each gene locus in the PR2 plan four quadrants distribution is not uniform, distribution in the fourth quadrant gene(the ratio of A3/(T3+G3)<0.5&G3/(G3+C3)>0.5) significantly more than the other three quadrants, indicating that the use of the third codon T nucleobase frequency is higher than that of A nucleobases, G nucleobases use frequency is higher than C nucleobasess. When RSUB is only affected by mutations, the frequencies of the four nucleobases should be used equally. Therefore, the results suggest that the codon usage pattern of the*C. panzhihuaensis panzhihuaensis* genome is influenced not only by mutations, but also by other factors such as selection pressure.

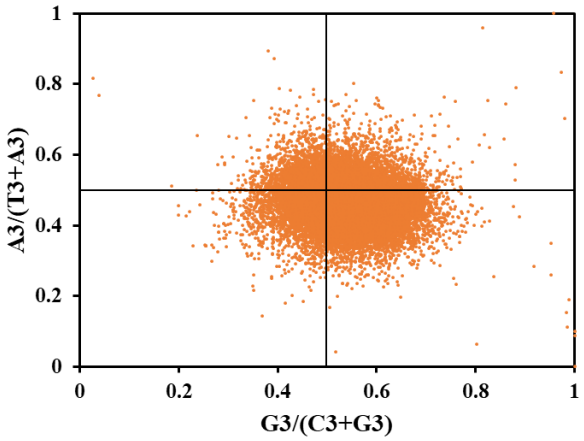

Figure 6. PR2-plot [A3/(A3+T3) against G3/(G3+C3)]

3.6. Effects of gene expression level

To analyze the effect of gene expression level on codon preference characteristics, the CAI value is an important index used to assess gene expression level. Consequently, this study calculated the correlation coefficient between the CAI value and gene indexes such as GC12, GC3s, GCall content, and ENC value. The results are shown in Figures 7 and Table 2,The CAI value and ENC

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 May 2022 doi:10.20944/preprints202205.0208.v1

value showed significant negative correlation (R2=-0.278, P < 0.01), demonstrating that gene expression level had significant influence on codon preference. In addition, CAI value showed a remarkable positive correlation with GC3 (r = 0.159, p < 0.001), GC12 (r = 0.059, p < 0.001) and negative correlation with GCall (r = 0.037, p < 0.001) content

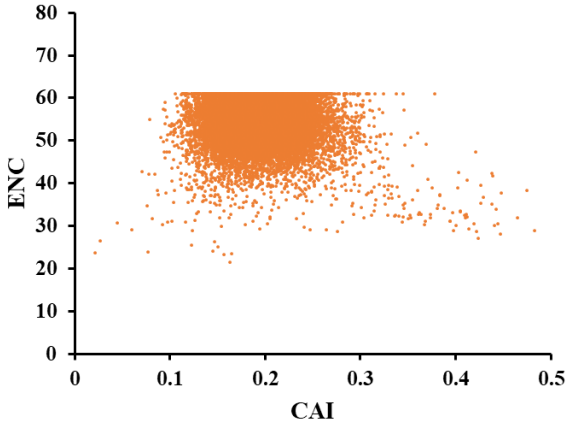

Figure 7 Plot of ENC versus gene expression level

3.7. Determination of optimal codons for*C. panzhihuaensis panzhihuaensis*

We carried out a two-way Chi-squared contingency test to compare the codon usage among different genes. The highly- and lowly- expressed data of the genes based on average RSCU values were list in the Table 3. The results showed that, as shown in Table 3, A total of 31 codons were found to have extremely significant differences (P < 0.01) between the high and low expression sample libraries. These codons were the optimal codons, and all the optimal codons ended in A/U except UUG of Leu and AGG of Arg. The result that the optimal codon prefers A/U nucleotide nucleobases endings is similar to the previous RSCU analysis.

Table 3 Optimal codons of*C. panzhihuaensis panzhihuaensis* genes based on the RSCU values

Amino acid	Codon	High expression		Low expression		Amino acid	Codon	High expression		Low expression	
		RSCU	Number	RSCU	Number			RSCU	Number	RSCU	Number
Phe	**UUU***	1.43	19172	0.6	5651	Ser	**UCU***	1.73	21146	0.85	5576
	UUC	0.57	7636	1.4	13212		UCC	0.54	6600	1.59	10381
Leu	**UUA***	0.98	12235	0.31	2280		**UCA***	1.67	20438	0.37	2385
	UUG*	1.61	20079	1.06	7797		UCG	0.22	2743	1.45	9483
	CUU*	1.59	19806	0.64	4709		**AGU***	1.18	14470	0.46	2988
	CUC	0.42	5210	1.51	11191		AGC	0.66	8090	1.28	8336
	CUA*	0.64	7906	0.23	1690	Pro	**CCU***	1.66	15273	0.74	5249
	CUG	0.75	9379	2.26	16672		CCC	0.43	3926	1.38	9753
Ile	**AUU***	1.61	21020	0.93	6499		**CCA***	1.72	15822	0.52	3708
	AUC	0.5	6578	1.49	10395		CCG	0.2	1844	1.35	9576
	AUA*	0.88	11543	0.57	3996	Thr	**ACU***	1.45	14175	0.52	3216
Val	**GUU***	1.66	21503	0.8	6737		ACC	0.48	4690	1.46	8935
	GUC	0.47	6150	1.15	9660		**ACA***	1.85	18091	0.47	2879
	GUA*	0.83	10748	0.38	3220		ACG	0.22	2195	1.55	9503
	GUG	1.03	13396	1.66	13938	Ala	**GCU***	1.55	22958	0.66	6745

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 May 2022 doi:10.20944/preprints202205.0208.v1

Tyr	**UAU***	1.44	12816	0.66	3910		GCC	0.48	7083	1.44	14743
	UAC	0.56	4998	1.34	8012		**GCA***	1.8	26641	0.69	7023
TER	UAA	0.96	470	0.88	422		GCG	0.17	2442	1.21	12334
	UGA	1.24	610	1.4	675	Cys	**UGU***	1.28	9670	0.55	2750
	UAG	0.8	393	0.72	348		UGC	0.72	5481	1.45	7334
His	**CAU***	1.53	15793	0.74	4434	Met	AUG	1	18706	1	10150
	CAC	0.47	4888	1.26	7624	Trp	UGG	1	9015	1	7475
Gln	**CAA***	1.11	20112	0.63	4966	Arg	**CGU***	0.79	5663	0.69	3944
	CAG	0.89	15964	1.37	10925		CGC	0.34	2459	1.14	6469
Asn	**AAU***	1.47	24524	0.86	7816		CGA	0.73	5231	0.74	4194
	AAC	0.53	8831	1.14	10346		CGG	0.52	3732	1.1	6229
Lys	AAA*	1.01	24239	0.7	7097		**AGA***	2.01	14467	1.08	6147
	AAG	0.99	23754	1.3	13096		**AGG***	1.61	11581	1.25	7078
Asp	**GAU***	1.51	33673	0.85	10237	Gly	**GGU***	1.23	16673	0.66	6339
	GAC	0.49	11067	1.15	13803		GGC	0.65	8787	1.31	12664
Glu	**GAA***	1.16	34865	0.81	10202		**GGA***	1.39	18913	0.96	9239
	GAG	0.84	25090	1.19	14985		GGG	0.73	9948	1.07	10364

Notes.

*Codons that occur significantly more often ($p < 0.01$).

The optimal codons are displayed in bold.

4 Discussion

Each organism forms a specific codon usage pattern in the long-term evolution process, and GC content is an important indicator of nucleobase composition in the genome, which is of great significance in the evolution of the genome. GC content usually reflects the strength of directional mutation. Although synonymous mutation of the third nucleobase of the codon cannot change the type of amino acids, it has always been considered an important feature to determine the type of amino acids, and GC3s are commonly used as an important indicator of codon bias[27,28]. In this research, it was discovered that the average GC content was close to GC3s content (both moderately less than 50%), indicating that the overall AU content in the C. panzhihuaensis genome was slightly higher than GC, and the codon tended to end in A/U slightly. Interestingly, this result is similar to GC3s content in gymnosperms such as *Ginkgo biloba* [31], *Gnetifer Hypothesis*[32], and most dicotyledons[32–35]. However, monocotyledons such as *Triticum Aestivum* [36],*Zea Mays*[37], *Hordeum Vulgare*[38], and *Oryza Sativa* [39] all showed high GC content and tended to end in G/C. RSCU analysis results reveal that similar to most higher plants [40], there is A or T use bias in the genome of C. panzhihuaensis. These results may imply that dicotyledons and gymnosperms are evolutionarily conserved relative to monocotyledons. In addition, the formation process of the codon usage pattern is usually affected by many factors, including mutation and selection. In this study, the neutral plot shows that GC12 is positively correlated with GC3s ($R2=0.292$, $P < 0.001$) (Figure 2). Enc-plot analysis indicated that most of the genes were distributed around the standard curve, while some genes were distributed far below the standard curve. Combined with Wright(1990), A method was proposed to distinguish selection pressure from mutation pressure. If the direction of mutation was the main cause of SCUB, G, C, A, and T should be evenly used in the codon of the A gene. However, as shown in the PR2-plot (Figure. 4), the frequency of use of the

 doi:10.20944/preprints202205.0208.v1

four nucleobases at the third position of the codon was distinct, which demonstrated that natural selection played a major role in the SCUB generation process of the C. panzhihuaensis genome, and mutations and other factors played an important role.

In this study, it was found that there were related factors between the nucleobases of *C. panzhihuaensis* genome sequence and SCUB. The correlation coefficients of GC12, CG3s, GCall, and Axis1 displayed extremely significant correlation (Table 2), indicating that the codon usage characteristics were considerably affected by the difference in nucleobases composition. In addition, it is very difficult to confirm the expression level of a gene in different tissues and developmental stages of differentiated multi-cellular eukaryotes. The codon adaptation index CAI has generally been used to calculate the gene expression level[41,42]. In this study, the correlation coefficient between the CAI value and the principal axis of the first vector showed a significant positive correlation ($R2=0.013$, $P < 0.05$), and the correlation analysis between the Axis and nucleobases and CAI values showed that the codon usage pattern of *C. panzhihuaensis* genome was mainly affected by the combined action of nucleobases difference and gene expression level, and the nucleobases difference was the dominant factor. Only a few plants have been sequenced, and it is a complex process to analyze the codon usage patterns and their influencing factors. Therefore, this study preliminarily revealed the codon usage characteristics of the *C. panzhihuaensis* genome, which will provide guidance for further studies on the phylogenetic evolution of *Cycad* and the function of related genes at the molecular level.

Under the combined action of mutation pressure and strong forward selection, it is easy to form a large number of optimal codons, while the combined action of mutation pressure and purification selection generally inhibits the formation of optimal codons[43]. In this study, a total of 31 optimal codons were screened by combining the results of high expression codon analysis and high-frequency codon analysis of *C. panzhihuaensis* genome, and most of the codons ended in U or A. At present, most of the optimal codons of higher plant genomes reported end in U or A, which may be correlated with the relative conservatism of genome evolution [44]. At the same time, the optimal codon and its number varied among different species, indicating that different species faced different evolutionary pressures during evolution.

5 Conclusions

This study on the genome level of *C. panzhihuaensis* codon usage pattern has carried on the comprehensive analysis of these results not only helps further understand *Cycad* gene expression with the relationship between the SCUB, and facilitates future cycad codon optimization in the process of genetic transformation, and also to explore the evolutionary *Cycad* and germplasm innovation is of great significance.

Author Contributions: Yongjian Luo, Ru Wang and Qing Li conceived and designed the project. Yongjian Luo analyzed the data. Songquan Song, Jun Liu and Zhijun Deng wrote the paper. All authors read and consented to the final version of the manuscript

Funding: This work was supported by the National Natural Science Foundation of China (31860073, 81303169, 31871716, 31371715), the Open Fund of Hubei Key Laboratory of Biologic Resources Protection and Utilization (PT012008),the Science and Technology Program of Guang- dong Province, China (2020B121201008, 2020B020209003, 2018B020202004), the

Science and Technology Program of Guangzhou (201909020001, 201807010114) and Agricultural Science and Technology Cooperation Project of Foshan (201908).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Excluded this statement.

Conflicts of Interest: The authors declare that they have no competing financial interests.

References

1. Ikemura, T. Codon Usage and TRNA Content in Unicellular and Multicellular Organisms. *Mol Biol Evol* **1985**, *2*, 13–34, doi:10.1093/oxfordjournals.molbev.a040335.
2. Parvathy, S.T.; Udayasuriyan, V.; Bhadana, V. Codon Usage Bias. *Mol Biol Rep* **2022**, *49*, 539–565, doi:10.1007/s11033-021-06749-4.
3. Duret, L.; Mouchiroud, D. Expression Pattern and, Surprisingly, Gene Length Shape Codon Usage in *Caenorhabditis* , *Drosophila* , and *Arabidopsis*. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 4482–4487, doi:10.1073/pnas.96.8.4482.
4. Srivastava, S.; Chanyal, S.; Dubey, A.; Tewari, A.; Taj, G. Patterns of Codon Usage Bias in WRKY Genes of Brassica Rapa and Arabidopsis Thaliana. *Journal of Agricultural Science* **2019**, *11*, 76, doi:10.5539/JAS.V11N4P76.
5. Nie, X.; Deng, P.; Feng, K.; Liu, P.; Du, X.; You, F.M.; Weining, S. Comparative Analysis of Codon Usage Patterns in Chloroplast Genomes of the Asteraceae Family. *Plant Mol Biol Rep* **2014**, *32*, 828–840, doi:10.1007/s11105-013-0691-z.
6. J Benjelloun; S Bouzroud; ZEA Triqui; QL Alami; A Guedira Warm Stratification Improves Embryos Development and Seed Germination of Cycas Revoluta. *Advances in Horticultural Science* **2021**, doi:10.36253/ahsc-9681.
7. Brenner, E.D.; Stevenson, D.W.; Twigg, R.W. Cycads: Evolutionary Innovations and the Role of Plant-Derived Neurotoxins. *Trends Plant Sci* **2003**, *8*, 446–452, doi:10.1016/S1360-1385(03)00190-0.
8. Nagalingum, N.S.; Marshall, C.R.; Quental, T.B.; Rai, H.S.; Little, D.P.; Mathews, S. Recent Synchronous Radiation of a Living Fossil. *Science* **2011**, *334*, 796–799, doi:10.1126/science.1209926.
9. Chen, C.J.; Liu, N. New Discoveries of Cycads and Advancement of Conservation of Cycads in China. *Botanical Review* **2004**, *70*, 93–100.
10. Kyoda, S.; Setoguchi, H. Phylogeography of Cycas Revoluta Thunb. (Cycadaceae) on the Ryukyu Islands: Very Low Genetic Diversity and Geographical Structure. *Plant Systematics and Evolution* **2010**, *288*, 177–189, doi:10.1007/s00606-010-0322-1.
11. Hassan, M.; Gomez, G.; Pallás, V.; Myrta, A.; Rysanek, P. Simultaneous Detection and Genetic Variability of Stone Fruit Viroids in the Czech Republic. *Eur J Plant Pathol* **2009**, *124*, 363–368, doi:10.1007/s10658-008-9420-0.
12. Ragasa, C.Y.; Ng, V.A.S.; Agoo, E.M.G.; Shen, C.-C. Chemical Constituents of Cycas Zambalensis. *Chem Nat Compd* **2016**, *52*, 136–138, doi:10.1007/s10600-016-1571-1.

13. Chemical Constituents of Cycas Sancti-Lasallei. *J App Pharm Sci* **2016**, doi:10.7324/JAPS.2015.54.S3.
14. Zheng, Y.; Liu, J.; Feng, X.; Gong, X. The Distribution, Diversity, and Conservation Status of *Cycas* in China. *Ecol Evol* **2017**, *7*, 3212–3224, doi:10.1002/ece3.2910.
15. Liu, Y.; Wang, S.; Li, L.; Yang, T.; Dong, S.; Wei, T.; Wu, S.; Liu, Y.; Gong, Y.; Feng, X.; et al. The Cycas Genome and the Early Evolution of Seed Plants. *Nat. Plants* **2022**, *8*, 389–401, doi:10.1038/s41477-022-01129-7.
16. Meier, N.; Meier, B.; Peter, S.; Wolfram, E. In-Silico UHPLC Method Optimization for Aglycones in the Herbal Laxatives Aloe Barbadensis Mill., Cassia Angustifolia Vahl Pods, Rhamnus Frangula L. Bark, Rhamnus Purshianus DC. Bark, and Rheum Palmatum L. Roots. *Molecules* **2017**, *22*, 1838, doi:10.3390/molecules22111838.
17. Zhang, W.-J.; Zhou, J.; Li, Z.-F.; Wang, L.; Gu, X.; Zhong, Y. Comparative Analysis of Codon Usage Patterns Among Mitochondrion, Chloroplast and Nuclear Genes in Triticum Aestivum L. *J Integrative Plant Biology* **2007**, *49*, 246–254, doi:10.1111/j.1744-7909.2007.00404.x.
18. Zhang, X.; Hu, Y.; Liu, M.; Lang, T. Optimization of Assembly Pipeline May Improve the Sequence of the Chloroplast Genome in Quercus Spinosa. *Scientific Reports* **2018**, *8*, doi:10.1038/s41598-018-27298-0.
19. Wu, Y.; Li, Z.; Zhao, D.; Tao, J. Comparative Analysis of Flower-Meristem-Identity Gene APETALA2 (AP2) Codon in Different Plant Species. *Journal of Integrative Agriculture* **2018**, *17*, 867–877, doi:10.1016/S2095-3119(17)61732-5.
20. Anue Mensah Raphael; Xueli Sun; Chunzhen Cheng; Zhongxiong Lai Analysis of Codon Usage Pattern of Banana Basic Secretory Protease Gene. *Plant Diseases and Pests* **2019**, *10*, 1-4+9, doi:10.19579/j.cnki.plant-d.p.2019.01.001.
21. Sharp, P.M.; Li, W.-H. The Codon Adaptation Index-a Measure of Directional Synonymous Codon Usage Bias, and Its Potential Applications. *Nucl Acids Res* **1987**, *15*, 1281–1295, doi:10.1093/nar/15.3.1281.
22. Duan, H.; Zhang, Q.; Wang, C.; Li, F.; Tian, F.; Lu, Y.; Hu, Y.; Yang, H.; Cui, G. 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. *PeerJ* **2021**, *9*, e10787, doi:10.7717/peerj.10787.
23. 丁锐; 胡兵; 宗小雁; 韩辰阳; 张丽杰; 陈旭辉 杓兰叶绿体基因组密码子偏好性分析. 林业科学研究 **2021**, *34*, 177–185, doi:10.13275/j.cnki.lykxyj.2021.005.021.
24. Sueoka, N. Near Homogeneity of PR2-Bias Fingerprints in the Human Genome and Their Implications in Phylogenetic Analyses. *Journal of Molecular Evolution* **2001**, *53*, 469–476, doi:10.1007/s002390010237.
25. Wright, F. Comparative Analysis of Flower-Meristem-Identity Gene APETALA2 (AP2) Codon in Different Plant Species. *Gene* **1990**, *87*, 23–29, doi:10.1016/0378-1119(90)90491-9.
26. Sueoka, N.; Kawanishi, Y. DNA G+C Content of the Third Codon Position and Codon Usage Biases of Human Genes. *Gene* **2000**, *261*, 53–62, doi:10.1016/S0378-1119(00)00480-7.
27. Liu, Q.; Feng, Y.; Xue, Q. Analysis of Factors Shaping Codon Usage in the Mitochondrion Genome of Oryza Sativa. *Mitochondrion* **2004**, *4*, 313–320,

doi:10.1016/j.mito.2004.06.003.

28. Liu, Q.; Xue, Q. Comparative Studies on Codon Usage Pattern of Chloroplasts and Their Host Nuclear Genes in Four Plant Species. *Journal of genetics* **2005**, *84*, 55–62, doi:10.1007/BF02715890.
29. Sablok, G.; Nayak, K.C.; Vazquez, F.; Tatarinova, T. Synonymous Codon Usage, GC3, and Evolutionary Patterns Across Plastomes of Three Pooid Model Species: Emerging Grass Genome Models for Monocots. *Molecular Biotechnology* **2011**, *49*, 116–128, doi:10.1007/s12033-011-9383-9.
30. Wright, F. The 'Effective Number of Codons' Used in a Gene. *Gene* **1990**, *87*, 23–29, doi:10.1016/0378-1119(90)90491-9.
31. He, B.; Dong, H.; Jiang, C.; Cao, F.; Tao, S.; Xu, L. Analysis of Codon Usage Patterns in Ginkgo Biloba Reveals Codon Usage Tendency from A/U-Ending to G/C-Ending. *Scientific Reports* **2016**, *6*, doi:10.1038/srep35927.
32. Majeed, A.; Kaur, H.; Kaur, A.; Das, S.; Joseph, J.; Bhardwaj, P. Codon Usage Pattern in Gnetales Evolved in Close Accordance with the Gnetifer Hypothesis. *Botanical Journal of the Linnean Society* **2021**, *196*, 423–436, doi:10.1093/botlinnean/boab006.
33. Li, N.; Sun, M.; Jiang, Z.; Shu, H.; Zhang, S. Genome-Wide Analysis of the Synonymous Codon Usage Patterns in Apple. *Journal of Integrative Agriculture* **2016**, *15*, 983–991, doi:10.1016/S2095-3119(16)61333-3.
34. Wang, L.; Xing, H.; Yuan, Y.; Wang, X.; Muhammad, S.; Tao, J.; Wei, F.; Zhang, G.; Song, X.; Sun, X. Genome-Wide Analysis of Codon Usage Bias in Four Sequenced Cotton Species. *PLoS ONE* **2018**, *13*, e0194372, doi:10.1371/journal.pone.0194372.
35. Huo, X.; Liu, S.; Li, Y.; Wei, H.; Gao, J.; Yan, Y.; Zhang, G.; Liu, M. Analysis of Synonymous Codon Usage of Transcriptome Database in *Rheum Palmatum*. *PeerJ* **2021**, *9*, e10450, doi:10.7717/peerj.10450.
36. Yang, C.; Zhao, Q.; Wang, Y.; Zhao, J.; Qiao, L.; Wu, B.; Yan, S.; Zheng, J.; Zheng, X. Comparative Analysis of Genomic and Transcriptome Sequences Reveals Divergent Patterns of Codon Bias in Wheat and Its Ancestor Species. *Front. Genet.* **2021**, *12*, 732432, doi:10.3389/fgene.2021.732432.
37. Liu, Q.; Xue, Q. Comparative Studies on Codon Usage Pattern of Chloroplasts and Their Host Nuclear Genes in Four Plant Species. *J Genet* **2005**, *84*, 55–62, doi:10.1007/BF02715890.
38. Mazumdar, P.; Binti Othman, R.; Mebus, K.; Ramakrishnan, N.; Ann Harikrishna, J. Codon Usage and Codon Pair Patterns in Non-Grass Monocot Genomes. *Ann Bot* **2017**, *120*, 893–909, doi:10.1093/aob/mcx112.
39. Liu, Q.; Feng, Y.; Zhao, X.; Dong, H.; Xue, Q. Synonymous Codon Usage Bias in Oryza Sativa. *Plant Science* **2004**, *167*, 101–105, doi:10.1016/J.PLANTSCI.2004.03.003.
40. MingZhao, S.; Fang, L.; JinPing, H.; KunBo, W. Analysis on codon usage of chloroplast genome of Gossypium hirsutum. *Scientia Agricultura Sinica* **2011**, *44*, 245–253.
41. Naya, H.; Romero, H.; Carels, N.; Zavala, A.; Musto, H. Translational Selection Shapes Codon Usage in the GC-Rich Genome of Chlamydomonas Reinhardtii. *FEBS Lett* **2001**, *501*, 127–130, doi:10.1016/s0014-5793(01)02644-8.
42. Gupta, S.K.; Bhattacharyya, T.K.; Ghosh, T.C. Synonymous Codon Usage in Lactococcus Lactis: Mutational Bias versus Translational Selection. *J Biomol Struct Dyn* **2004**, *21*, 527–

 doi:10.20944/preprints202205.0208.v1

536, doi:10.1080/07391102.2004.10506946.

43. Hershberg, R.; Petrov, D.A. Selection on Codon Bias. *Annu Rev Genet* **2008**, *42*, 287–299, doi:10.1146/annurev.genet.42.110807.091442.
44. Duret, L. Evolution of Synonymous Codon Usage in Metazoans. *Current Opinion in Genetics & Development* **2002**, *12*, 640–649, doi:10.1016/S0959-437X(02)00353-2.