INTERNATIONAL JOURNAL OF AGRICULTURE & BIOLOGY
ISSN Print: 1560–8530; ISSN Online: 1814–9596
20F–065/202x/00–0–000–000
DOI: 10.17957/IJAB/15.1487
http://www.fspublishers.org

Full Length Article

Complete Chloroplast Genome Sequences from Yellowhorn (*Xanthoceras sorbifolia*) and Evolution Analysis Based on Codon Usage Bias

Xiaonong Guo[1,2,3*], Yaling Wang[2] and Suomin Wang[1]
[1]*State Key Laboratory of Grassland Agro-ecosystems, College of Pastoral Agriculture Science and Technology, Lanzhou University, Lanzhou 730020, PR, China*
[2]*Key Laboratory of Biotechnology and Bioengineering of State Ethnic Affairs Commission，Biomedical Research Center, Northwest Minzu University, Lanzhou 730030, PR, China*
[3]*Life Science and Engineering College, Northwest Minzu University, Lanzhou 730030, PR, China*
[*]For correspondence: gxnwww@126.com
Received 24 March 2020; Accepted 18 April 2020; Published _____

Abstract

Yellowhorn (*Xanthoceras sorbifolia* Bunge) is an economically important tree, and possesses higher genetic diversity. To explore the factors affecting the use of synonymous codons, and also analyze the relationship of codon usage patterns and evolutionary factors in yellowhorn, the complete chloroplast (cp) genome was determined, and the codon usage bias was analyzed. The evolutionary forces of cp genome in yellowhorn were inferred. The size of yellowhorn cp genome was 159,474 bp, which comprised of 4 regions. A total of 48 genes were selected to analyze the codon usage, the codon usage bias in yellowhorn cp genome was weak and codons preferred A/T ending. The results of Principal component analysis (PCA) and Corresponding analysis (COA) indicated that base composition (such as GC content) for mutation bias might affect the codon bias. Furthermore, combined with the analysis of ENC-plot, parity rule 2 (PR2) plot, and neutrality plot, natural selection might dominate the codon usage bias in yellowhorn cp genome except the notable mutation pressure. Our results will help understand the genetic architecture and mechanisms in yellowhorn, and also contribute to enriching genetic resources and conservation of endemic yellowhorn species. © 2020 Friends Science Publishers

Keywords: Yellowhorn; Chloroplast genome; Synonymous codon usage; Evolutionary forces

Introduction

Yellowhorn (*Xanthoceras sorbifolia* Bunge), belonging to family Sapindaceae, is a deciduous tree species distributed naturally on hills and slopes in Northern China (Wang *et al*. 2019a). As an economically important tree, the oil content of seed (50–68% of kernel) is high, accompanied with high content unsaturated fatty acid (85–93%), which is mainly comprised by linoleic acid, oleic acid, and nervonic acid content (Ji *et al.* 2017). Moreover, yellowhorn exhibits strong stress resistance to cold (even below −40ºC), salinity, and drought, implying important ecological value (Ruan *et al.* 2017). Yellowhorn can be used to treat rheumatism, gout, and children enuresis. Besides, triterpenoid saponins and barringenol-like triterpenoids extracted from different yellowhorn tissues have antitumor and anti-inflammatory effects to treat Alzheimer disease (Ding *et al.* 2019).

Yellowhorn is an andromonoecious plant, evident differences can be found in the morphological and physiological indicators of flower and fruit, including size, color, yield and oil content, after years of natural hybridization and selection (Bi and Guan 2014). Yellowhorn possesses higher genetic diversity, which exhibits abundant phenotypic variation even though in a small range under the same management practices (Ruan *et al.* 2017). Unlike other plant species, yellowhorn can survive for hundreds of years (Wang *et al.* 2019b). Thus, evolution analyses can expand the genetic engineering for yellowhorn resources and help understanding the genetic diversity based on the geographical distribution.

Codons are the basic genetic codes of mRNA, the degeneracy of which exists in the process of encoding amino acids in all living organisms (Taylor and Coates 1989). The pattern of synonymous codon usage is not random, exhibit usage bias in different genomes (Sharp and Cowe 1991). Codon bias is unique to a given organism, and is influenced by a series of factors, such as GC content, gene lengths, gene expression levels and so on (Plotkin and Kudla 2010). The main two evolutionary forces of natural selection and mutation bias can be reflected by codon usage bias (Akashi 1994; Hershberg and Petrov 2008). Thus, codon usage bias can provide clues on plant species evolution.

To cite this paper: Guo X, Y Wang, S Wang, (202x). Complete chloroplast genome sequences from yellowhorn (*Xanthoceras sorbifolia*) and evolution analysis based on codon usage bias. *Intl J Agric Biol* 00: 000–000

Yellowhorn is an economically important tree, exhibits the strong stress resistance (Wang *et al.* 2017). High-quality yellowhorn genome database has been generated recently by Liang *et al.* (2019). The comparative *de novo* transcriptome analysis of yellowhorn was conducted by Zhou and Zheng (2015). Although the genetic resources of yellowhorn are increasing, it is very important to analyze the synonymous codons usage, which is currently unknown. The cp genome data can provide molecular phylogenetic information for developing commercially important yellowhorn species. In the present study, we obtained the complete cp genome sequences in yellowhorn, which comprised by a pair of IR, LSC, and SSC regions. We identified complete characteristics of the yellowhorn chloroplast (cp) genome, focusing on the codon usage bias by using multivariate statistical analysis, and analyzed the evolutionary forces preference. Our results will provide the information to help understanding the genetic architecture and mechanisms in yellowhorn, and also contribute to enriching genetic resources and conservation of endemic yellowhorn species.

Materials and Methods

Experimental material

Yellowhorn was planted in Wenhuagong in the Loess Plateau Area (36°36′N，103°48′E), Lanzhou city, Gansu province, China. Fresh leaves were collected on August 29th 2019, and then kept at −80°C after immediately frozen.

DNA extraction, sequencing and assembly

Genomic DNA was extracted by using the method of Li *et al*. (2018). The integrity and quality of DNA was validated by a spectrophotometer (OD-1000，Shanghai, China). With reference to the NEBNext® Ultra™ DNA Library Prep Kit for Illumina® instruction, the library with 250 bp length was constructed and sequenced on an Illumina NovaSeq platform (Benagen Tech Solution Co., Ltd, Wuhan, China). After the Illumina PCR adapter reads, low-quality reads and reads of more than 5% unknown nucleotide "Ns" were filtered from the paired-end raw reads in the quality control step. All good-quality paired clean reads were obtained using SOAPnuke software, version: 1.3.0 (Chen *et al*. 2017). After performing bidirectional iterative derivation of the assembled reads by NOVOPlasty (k-mer = 39) (Dierckxsens *et al*. 2016), the whole circular genome sequence was obtained. All circled sequences were searched by BLASTN (version: BLAST 2.2.30+, E-value $\leq 1^{e-5}$) against the reference database (Goodwin *et al*. 2015).

Phylogenetic analysis

The total of 31 complete cp genome data from different plant species were downloaded from the NCBI database and the sequences alignment of which was initially conducted using MAFFT (Katoh *et al*. 2002). The phylogenetic tree was generated by MEGA-X (Kumar *et al.* 2018).

Codon usage bias

Based on the cp genome data of yellowhorn, filtering the repeated sequences and the sequences length less than 300 bp, 48 sequences with the CDSs were retained to do the analysis of codon usage bias (Comeron and Aguadé 1998). The important indicators were performed by using codon W software version 1.3 (https://sourceforge.net/projects/codonw/), including RSCU (the relative synonymous codon usage value), ENC (the effective number of codons), CAI (the codon adaptation index), GC (G + C content of the gene), GC_{3s} (the frequency of the nucleotides G + C at the 3rd of synonymous codons), and the base compositions (A_{3s}, T_{3s}, G_{3s}, and C_{3s}) (Puigbò *et al*. 2008). The G+C content at the 1st, 2 nd, 3rd of codons (GC_1, GC_2, GC_3) and GC_{12} (the average GC content of 1st and 2nd) were calculated by Cusp function from EMBOSS (http://imed.med.ucm.es/EMBOSS/) (Rice *et al*. 2000).

Identification of the optimal codon

Using ENC values as preference standard, 48 sequences of yellowhorn were ordered, and 5% high bias dataset and 5% low bias dataset were selected. ΔRSCU of the codon was calculated by the RSCU value of each codon with high bias minus the RSCU value with low bias (Sharp and Li 1987). Finally, the optimal codon of the gene was speculated by the codon possessing the highest and largest ΔRSCU.

Multivariate statistical analysis

The distribution of all the genes under a 59 vector space according to the RSCU values was analyzed by Principal component analysis (PCA). The data with different axes were obtained, and the axes consistent with the most important factors which held important implications of codon usage variation were revealed (Wold *et al.* 1987). Corresponding analysis (COA) was used to compare two or more categories of variable data, and provide the visual results of the major changes in trend of codon usage and genes (Perriere and Thioulouse 2002). ENC-plot mapping analysis was used to identify the key factors affecting the codon usage bias. The ENC plot of the ENC values against the GC_{3S} values was drawn by EXCEL 2016. The ideal relationship of ENC and GC_{3s} can be observed from the standard curve.

Parity rule 2 (PR2) plot mapping analysis was constructed to show the relationship of the values $A_3/(A_3 + T_3)$ and $G_3/(G_3 + C_3)$, and the data were distributed into four quadrants in a scatter diagram (Sueoka 1999). Analysis of the relationship between GC_{12} and GC_3 values of all genes was performed by using neutrality plot mapping. In the neutral graph, the value of GC_{12} is used as vertical

coordinate, and the value of GC_3 is used as horizontal axis (Wei *et al*. 2014). The correlation analysis among many important indices was calculated using SPSS 16.0 software with the Spearman's test (two-tailed).

Results

Features of yellowhorn cp genome

A 34.6 million raw reads were obtained and 3.46 Gb clean reads were selected. The data was deposited in the Genbank database (Accession number: MN608158). The size of yellowhorn cp genome was 159,474 bp, which comprised of a pair of 54,496 bp IR (inverted repeat, IRa and IRb), 86,298 bp LSC (the large single copy), and 18,680 bp SSC (small single copy) regions (Fig. 1). The positions of the 114 genes identified in the yellowhorn cp genome are shown in Fig. 1. Major portion (66.7%) of the 78 genes were protein-coding genes, whereas, the RNA-coding genes comprised 33.3% (including 31 tRNA-coding genes and 8 rRNA-coding genes). The overall A + T content was 62.3%. The A + T content of the IR regions was 57.4%, whereas those of LSC and SSC regions were 64.0% and 68.3%, respectively.

Phylogenetic analysis

Phylogenetic analysis was completed on an alignment of 31 complete cp genome data from 31 plant species (Fig. 2). The results indicated that yellowhorn clustered together with the cp genome of *X. sorbifolium*, which was the homotypic synonym of *X. sorbifolia*, implying that no significant evolutionary difference was found for yellowhorn with different ages and growing regions. Besides, high homology between yellowhorn and *Acer buergerianum* was observed, which were belonged to the same family (Sapindaceae).

The codon usage pattern of yellowhorn cp genome

The average content of GC, GC_1, GC_2 and GC_3 of the cp genome of yellowhorn was 39, 39, 29 and 49%, respectively. The frequencies of A_{3s}, T_{3s}, G_{3s}, C_{3s}, and GC_{3s} were 43, 46, 17, 17 and 26%, respectively. The amino acids number of 48 genes was between 102-2297 with an average of 416. The ENC values ranged from 40.7 to 56.8 and the average was 48.8. All CAI values of these 48 sequences ranged from 0.1 to 0.3, which were far less than 1 (Table 1).

The statistic description of each codon in yellowhorn cp genome was shown, 18 high frequency used synonymous codons were observed, all the RSCU values of these 18 codons were more than 1.2, which preferred ending with T or A (T: 13 ones, A: 5 ones) (Table 2). 24 codons were identified as the high expressed codons (Table 3). 9 codons with high frequency codons as well as high expressed codons including GCT, GGT, ATT, AAA, CCT, CAA, AGA, TCT, and ACT were characterized as the optimal codons, of which, 6 were ending with T, and 3 were ending with A.

PCA analysis

Forty eight genes of yellowhorn cp genome were performed to do the PCA analysis, and were distributed in 47 dimensional axes. The contribution of 40 axes was shown in Fig. 3, the genes variations from Axis 1 to Axis 4 accounted for 35.48% of the total axes variation. Axis 1 and Axis 2 explained 10.26 and 9.85% of the total variation, meanwhile, Axis 3 and Axis 4 explained 8.04% and 7.33% of that, suggesting that the total of four axes were important for the codon usage bias.

COA analysis

After COA analysis, the location of codons ending with different bases was drawn by different color points between Axis 1 and Axis 2 (Fig. 4a). No significant pattern between codons with different bases ends and the two axes was found, although the codons with A/T ends were more tightly classified than those with G/C ends (Fig. 4a). Moreover, the location of different gene types was also drawn by different color points between Axis 1 and Axis 2 (Fig. 4b). The gene *rbcL* from Rubisco large subunit, genes of Cytochrome b/f complex and RNA polymerase were located in only one quadrant. The points of ClpP, matK, photosystem I, photosystem II, and hypothetical chloroplast reading frames (*ycf*) were distributed in two different quadrants. However, genes of ATP synthase, NADH dehydrogenase, and, ribosomal proteins (LSU), ribosomal proteins (SSU), and other genes distributed discretely. The results of genes distribution suggested that different classes of genes possessed different codon usage patterns.

In order to analyze the relationship of the important indices to the four main axes, correlation analysis was conducted to analyze the crucial factors influencing codon usage bias (Table 4). GC content showed the significant positive correlation with Axis 1 ($r = 0.358$, $p < 0.05$), meanwhile, a significant positive correlation between CAI value and Axis 1 ($r = 0.491$, $p < 0.01$) was also found, suggesting that the gene expression level might have the effect on the codon bias except for the nucleotide content (such as the GC content) of the genes. In addition, GC_{3s} exhibited the significant positive correlation with Axis 2 ($r = 0.299$, $p < 0.05$).

ENC plot analysis

The relationship of GC_{3S} and ENC value of genes was analyzed and the distribution trend was shown in Fig. 5a. Some genes, for example, 3 members of *ycf* (*ycf1*, *ycf2*, and *ycf4*), all of the 4 members belonged to RNA polymerase (*rpoA*, *rpoB*, *rpoC1*, and *rpoC2*) were located on or close to the curve. However, most of the genes lied away from the standard curve, accompanied with a relative concentrate distribution. In addition, the correlation analysis of ENC values and GC_3 values showed the extreme positive

Table 1: Indices of codon usage in cp genome of yellowhorn. GC: G+C content of the gene, GC_3: The G+C content at the 3_{rd} of codons, GC_{3S}: the frequency of the nucleotides G+C at the 3_{rd} of synonymous codons, CAI: the codon adaptation index, ENC: the effective number of codons.

Gene	GC	GC_3	GC_{3S}	CAI	ENC	Gene	GC	GC_3	GC_{3S}	CAI	ENC
cemA	0.34	33	0.30	0.17	56.84	*ycf2*	0.38	37	0.35	0.16	53.31
clpP	0.43	32	0.27	0.17	55.52	*psbA*	0.42	33	0.29	0.30	41.55
accD	0.36	29	0.26	0.20	47.71	*psbC*	0.45	35	0.30	0.19	48.75
atpA	0.41	27	0.26	0.19	48.47	*psbB*	0.43	29	0.25	0.19	46.52
rp122	0.35	27	0.25	0.16	51.41	*ndhC*	0.36	29	0.22	0.22	47.10
atpI	0.38	27	0.24	0.17	45.12	*matK*	0.35	30	0.28	0.16	50.77
rps4	0.39	27	0.26	0.15	54.13	*psbD*	0.42	31	0.26	0.24	44.99
atpE	0.41	32	0.30	0.17	51.78	*rpoA*	0.35	27	0.24	0.17	49.20
ndhK	0.39	28	0.26	0.17	53.46	*rp12*	0.44	33	0.31	0.14	54.67
rpoB	0.40	32	0.29	0.15	51.70	*atpB*	0.44	32	0.30	0.20	47.57
psaA	0.43	32	0.28	0.20	50.08	*rps18*	0.35	26	0.23	0.11	37.50
ndhD	0.37	30	0.26	0.14	51.84	*ccsA*	0.37	32	0.28	0.13	55.50
petA	0.40	30	0.30	0.18	50.92	*rps3*	0.35	21	0.19	0.15	47.25
ndhB	0.38	31	0.28	0.16	48.08	*rps14*	0.41	33	0.30	0.15	41.17
petD	0.40	29	0.27	0.17	46.61	*atpF*	0.39	31	0.30	0.16	44.52
ndhA	0.35	22	0.19	0.13	42.24	*psaB*	0.41	33	0.28	0.18	50.46
ndhE	0.34	27	0.23	0.15	48.12	*rsoC2*	0.38	31	0.29	0.15	51.39
ndhJ	0.40	32	0.28	0.17	55.84	*rps7*	0.41	25	0.23	0.19	43.20
ndhI	0.35	23	0.20	0.23	45.13	*rpoC1*	0.39	27	0.25	0.15	48.91
ndhG	0.36	28	0.25	0.14	48.60	*rps8*	0.35	25	0.21	0.10	40.73
ndhH	0.39	29	0.24	0.17	50.12	*rp114*	0.42	31	0.30	0.16	53.57
ycf4	0.37	30	0.25	0.18	48.02	*rp116*	0.43	26	0.20	0.16	41.60
ycf3	0.40	34	0.32	0.15	54.00	*rbcL*	0.45	33	0.30	0.24	50.19
ycf1	0.31	26	0.23	0.17	47.98	Average	0.39	29.46	0.26	0.17	48.77
rp120	0.36	27	0.23	0.10	46.66						

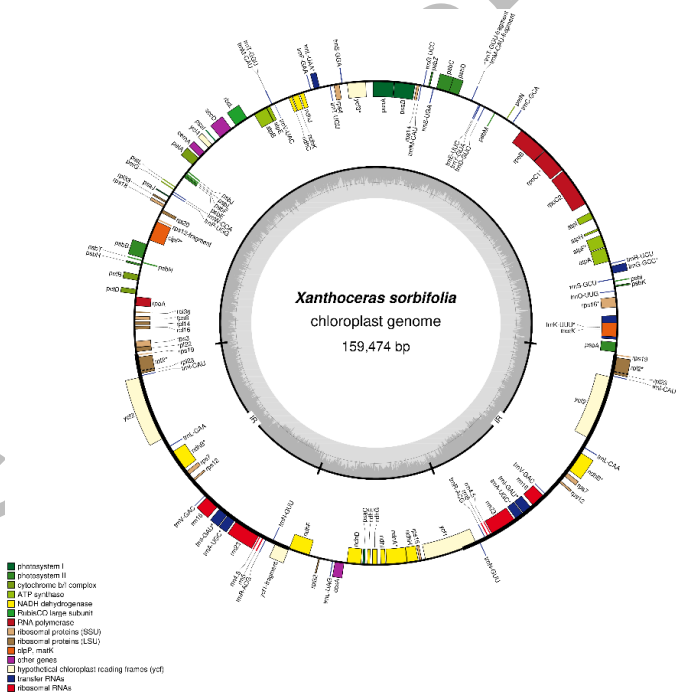

Fig. 1: Circular yellowhorn cp genome map. Genes with different boxes inside or outside the circle represent the direction of transcription. Different colors indicate the gene functional group

correlation ($r = 0.478$, $p < 0.01$), implying that the third position of codons might affect the codon usage bias.

PR2-plot mapping analysis

Using PR2 plot mapping analysis, the points in our plot located among 0.24 to 0.71 on $A_3/(A_3 + T_3)$, and 0.35 to 0.58 $G_3/(G_3 + C_3)$, indicating relative lower bias toward either A_3/T_3 or G_3/C_3 in yellowhorn. Furthermore, it was clearly shown that the genes distributed unevenly in the four quadrants, 14 and 17 genes were located in the third (in which the ratio of $A_3/(A_3 + T_3)$ and $G_3/(G_3 + C_3) < 0.5$) and

Table 2: Codon usage in yellowhorn cp genome. The preferentially used codons (RSCU > 1) are in bold. RSCU: relative synonymous codon usage

Amino acid	Codon	Number	RSCU	Amino acid	Codon	Number	RSCU
Ala(A)	**GCT**	**474**	**1.75**	Asn(N)	**AAT**	**726**	**1.55**
	GCC	178	0.66		AAC	210	0.45
	GCA	295	1.09	Pro(P)	**CCT**	**312**	**1.55**
	GCG	139	0.51		CCC	160	0.79
Cys(C)	**TGT**	**155**	**1.44**		CCA	220	1.09
	TGC	60	0.56		CCG	115	0.57
Asp(D)	**GAT**	**620**	**1.58**	Gln(Q)	**CAA**	**537**	**1.53**
	GAC	167	0.42		CAG	166	0.47
Glu(E)	GAA	810	1.5	Arg(R)	**CGT**	**241**	**1.24**
	GAG	271	0.5		CGC	98	0.5
Phe(F)	**TTT**	**709**	**1.3**		CGA	272	1.4
	TTC	382	0.7		CGG	91	0.47
Gly(G)	**GGT**	**466**	**1.34**		**AGA**	**336**	**1.73**
	GGC	138	0.4		AGG	129	0.66
	GGA	543	1.56	Ser(S)	**TCT**	**372**	**1.56**
	GGG	245	0.7		TCC	234	0.98
His(H)	**CAT**	**340**	**1.48**		TCA	290	1.22
	CAC	120	0.52		TCG	138	0.58
Ile(I)	**ATT**	**805**	**1.45**		AGT	292	1.23
	ATC	340	0.61		AGC	104	0.44
	ATA	516	0.93	Sec(T)	**ACT**	**376**	**1.54**
Lys(K)	**AAA**	**794**	**1.51**		ACC	195	0.8
	AAG	257	0.49		ACA	294	1.21
Leu(L)	**TTA**	**629**	**1.85**		ACG	109	0.45
	TTG	414	1.22	Val(V)	**GTT**	**402**	**1.52**
	CTT	429	1.26		GTC	115	0.44
	CTC	134	0.39		GTA	398	1.51
	CTA	288	0.85		GTG	142	0.54
	CTG	150	0.44	Trp(W)	TGG	355	1
Met(M)	ATG	419	1	Tyr(Y)	**TAT**	**576**	**1.63**
					TAC	131	0.37

fourth (in which the ratio of $A_3/(A_3 + T_3) < 0.5$ and $G_3/(G_3 + C_3) > 0.5$) quadrant, individually, while only 8 and 9 genes were located in the first and second quadrant (Fig. 5b). The results of Pearson correlation analysis indicated that no significant correlation (r = 0.155) of $A_3/(A_3 + T_3)$ and $G_3/(G_3 + C_3)$ was found. The results above showed that the genes in yellowhorn had a slight yet noticeable preference for T at the third position of the codon. Thus, the balance between A/T and G/C in the yellowhorn was disrupted.

Neutrality plot analysis

From the neutrality plot, the relationship of GC_{12} and GC_3 was analyzed, and the change degree of natural selection and mutation pressure was estimated (Fig. 5c). Genes of *ycf2* and *cemA* located around the effected curve, the remaining genes were up the standard curve. Using Pearson correlation analysis, weak correlation of all coding genes between GC_{12} and GC_3 was found (r = 0.261).

Discussion

The relative conservative natures of cp genomes with both structure and gene content were found in many plant species, for example, Korean ginseng (Kim and Lee 2004), *Arabidopsis thaliana* (Sato *et al*. 1999), rice (Wang and Hickey 2007), *Lotus japonicus* (Kato *et al*. 2000) and so on, except for some plants (i.e. alfalfa) with the extreme contraction or loss of IR regions (Tao *et al*. 2017).

The codon use probability differs in the process of protein synthesis (Morton 1999; Ghosh *et al.* 2000). The method of a gene formation by using specific synonymous codons is useful for the evolutionary pattern of natural selection or mutation selection (Chen *et al*. 2011). In our study, we analyzed the codon usage bias of 48 genes in yellowhorn cp genome, and many important indices were calculated. ENC is an important indicator to reflect the preference degree of unequal use of synonymous codons (Wright 1990). The value of ENC less than 35 means strong codon preference, otherwise, weak codon preference will occur. The value of ENC in our study implied weak preference of synonymous usage. The codons of genes were rich in A/T, especially, 9 optimal codons with high frequency and high expression were all ended with A/T, implying the codons ending with C/G were lacking bias in the yellowhorn cp genome. Our results were consistent with other different plants, such as *Porphyra umbilicalis*, *Nuphar*, *Oncidium gower ramsey*, *Ranunculus*, and so on (Raubeson *et al*. 2007; Chen *et al*. 2011; Li *et al*. 2019). Therefore, a strong A/T bias of synonymous codon usage is universal in plant chloroplast genomes. Similar patterns across these plant species indicated that codon usage could

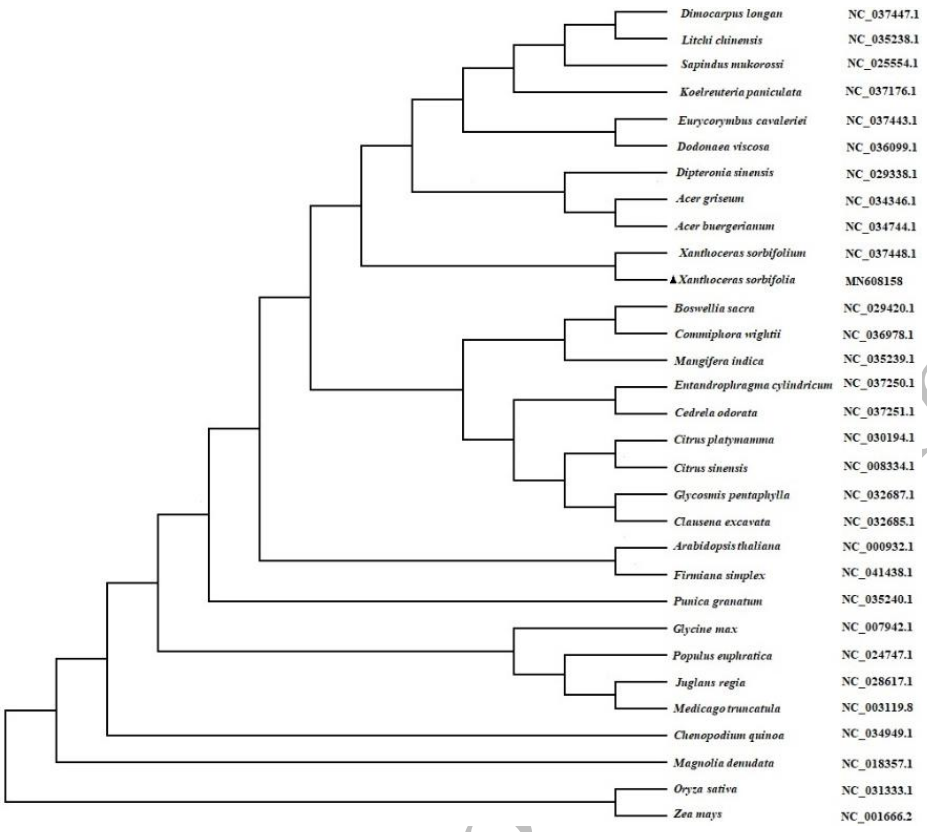

Fig. 2: Phylogenetic tree based on the cp genome data from 31 different plants. The sequence data of these plants were downloaded from NCBI database, and the accession numbers were shown on the tree. ▲ represents *X. sorbifolia* used in this study

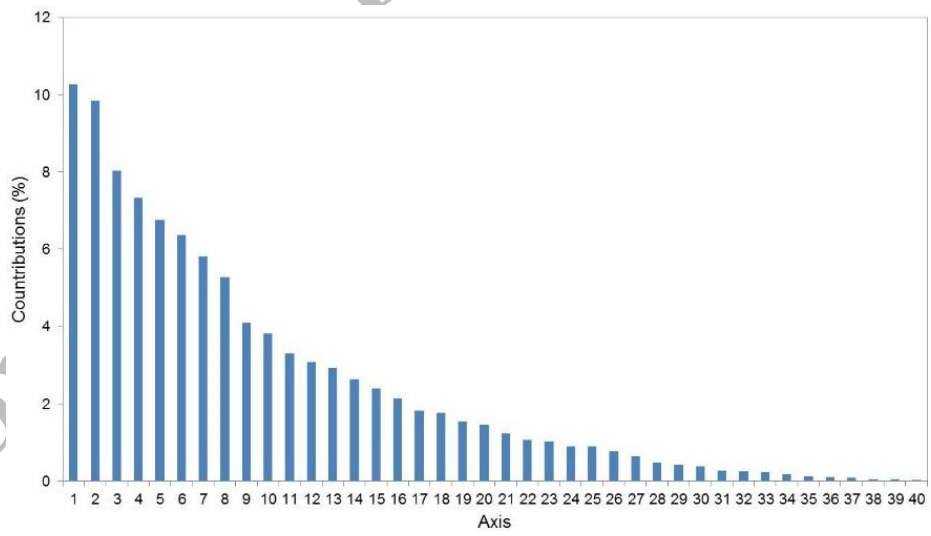

Fig. 3: Contributions of 40 axes from a principal component analysis (PCA) are shown

be regulated by universal biological factors in the long period evolution, for example, nucleotide composition.

We applied multivariate statistical analysis to explore the nucleotide composition effect on codon usage in yellowhorn. From the PCA analysis, Axis 1 and Axis 2 only accounted 20.11% variation of the total variation. After COA analysis, the codons ended with A/T or C/G did not show any pattern related to the Axis 1 and Axis 2, showing that A/T-endings were not crucial for the variation of codon usage bias. Usually, the base change from the third position

Table 3: The codons statistics with high and low expression genes of the yellowhorn cp genome. Codons with high expression level were shown with asterisk

Amino acid	Codon	High expressed gene		Low expressed gene		ΔRSCU
		Frequency	RSCU	Frequency	RSCU	
Ala (A)	GCU*	27	2.70	10	1.54	1.16
	GCC	4	0.40	6	0.92	-0.52
	GCA	8	0.80	5	0.77	0.03
	GCG	1	0.10	5	0.77	-0.67
Cys (C)	UGU	2	1.33	2	2.00	-0.67
	UGC*	1	0.67	0	0.00	0.67
Asp (D)	GAU	8	1.23	7	2.00	-0.77
	GAC*	5	0.77	0	0.00	0.77
Glu (E)	GAA	20	1.48	14	1.47	0.01
	GAG	7	0.52	5	0.53	-0.01
Phe (F)	UUU	10	0.71	11	1.47	-0.76
	UUC*	18	1.29	4	0.53	0.76
Gly (G)	GGC*	6	0.57	1	0.29	0.28
	GGA	10	0.95	9	2.57	-1.62
	GGG	1	0.10	1	0.29	-0.19
His (H)	CAU	5	0.91	4	2.00	-1.09
	CAC*	6	1.09	0	0.00	1.09
Ile (I)	AUU *	26	1.63	12	1.24	0.39
	AUC	11	0.69	9	0.93	-0.24
	AUA	11	0.69	8	0.83	-0.14
Lys (K)	AAA*	6	1.71	6	1.50	0.21
	AAG	1	0.29	2	0.50	-0.21
Leu (L)	UUA	13	1.70	7	1.83	-0.13
	UUG	9	1.17	5	1.30	-0.13
	CUU	9	1.17	5	1.30	-0.13
	CUC	2	0.26	2	0.52	-0.26
	CUA *	12	1.57	3	0.78	0.79
	CUG	1	0.13	1	0.26	-0.13
Met (M)	AUG	17	1.00	8	1.00	0.00
Asn (N)	AAU	18	1.16	16	1.60	-0.44
	AAC*	13	0.84	4	0.40	0.44
Pro (P)	CCU*	12	2.82	1	0.67	2.15
	CCC	0	0.00	3	2.00	-2.00
	CCA*	4	0.94	0	0.00	0.94
	CCG	1	0.24	2	1.33	-1.09
Gln (Q)	CAA*	7	1.75	6	1.20	0.55
	CAG	1	0.25	4	0.80	-0.55
Arg (R)	CGU	6	1.06	3	1.06	0.00
	CGC*	6	1.06	0	0.00	1.06
	CGA	10	1.76	6	2.12	-0.36
	CGG	1	0.18	5	1.76	-1.58
	AGA*	7	1.24	3	1.06	0.18
	AGG*	4	0.71	0	0.00	0.71
	GGU*	25	2.38	3	0.86	1.52
Ser (S)	UCU*	17	3.00	2	0.63	2.37
	UCC*	5	0.88	2	0.63	0.25
	UCA	1	0.18	7	2.21	-2.03
	UCG	0	0.00	3	0.95	-0.95
	AGU	7	1.24	4	1.26	-0.02
	AGC*	4	0.71	1	0.32	0.39
Thr (T)	ACU*	13	2.00	5	1.54	0.46
	ACC*	9	1.38	2	0.62	0.76
	ACA	3	0.46	5	1.54	-1.08
	ACG	1	0.15	1	0.31	-0.16
Val (V)	GUU	12	1.78	5	1.82	-0.04
	GUC	0	0.00	1	0.36	-0.36
	GUA*	15	2.22	2	0.73	1.49
	GUG	0	0.00	3	1.09	-1.09
Trp (W)	UGG	11	1.00	3	1.00	0.00
Tyr (Y)	UAU	11	1.29	14	1.75	-0.46
	UAC*	6	0.71	2	0.25	0.46

of codon cannot cause changes in the coding amino acids, for the less selection pressure (Guan *et al*. 2018). So it is very important to analyze the third position base composition of codon. GC content was positively correlated with Axis 1, also GC_{3s} was positively correlated with Axis 2; all these indicated that the base composition (such as GC content) might play a role on the codon bias.

The observed codon usage bias of genes is controlled by mutation pressure and selection (Wei *et al*. 2014). The ENC plot mapping analysis is helpful to understand the potential evolutionary factor. If codon usage of a particular gene is random, it will fall on or just below the standard curve (Raubeson *et al*. 2007). Otherwise, the genes far below the curve may be influenced by many factors, for example, GC bias of mutation pressure, and selection for codons ending in G/C. In our study, most of the genes lied below the curve, revealing a possibility that some factors (for example, natural selection) influenced codon bias to a certain extent except for mutation bias. Interestingly, the evolutionary factors in the formation of codon usage often changed in different gene classes. Mukhopadhyay *et al*. (2008) compared the codon usage differences between rice and Arabidopsis. They mainly focused on the two types of genes, and found that selective constraint of housekeeping genes were stronger than tissue-specific genes. In our study, most of the genes from *ycf* and RNA polymerase lied on or close to the curve, showing that the codon bias of them was only or mainly affected by mutation pressure. Nevertheless, most discrete distributed genes implied that they might be subject to different evolutionary factors.

It is noted that the relative effects of the two main evolutionary forces cannot be explained simply by looking at the ENC plot analysis (Liu *et al*. 2010). PR2 plot analysis showed that the location of different ending bases was asymmetric and preferred T-ending codons. It seemed that preferred codons undergo natural selection over long-term evolution, which was largely supported by neutrality plot mapping analysis. Neutrality plot mapping analysis is effective to learn the relationships of GC_{12} and GC_3. In the neutral graph of yellowhorn, no significant correlation of GC_3 and GC_{12} was found, suggesting strong difference between them. The results indicated that natural selection might be the most important factor affecting the codon usage bias of yellowhorn. Combined with the ENC plot, PR2 plot and neutrality plot, the results suggested that natural selection dominated the codon usage bias in yellowhorn cp genome.

Conclusion

The synonymous codon usage bias in yellowhorn cp genome was weak, and codons preferred A/T ending. Except the notable mutation pressure effects, majority of genetic evolution in yellowhorn was driven by natural selection.

Acknowledgements

This work was supported by the National Natural Science

Table 4: Correlation coefficients of the indices influencing codon bias in yellowhorn cp genome. Asterisk represents positive correlation ($P < 0.05$), **represents significant positive correlation ($P < 0.01$). GC: G+C content of the gene ENC: the effective number of codons, CAI: the codon adaptation index, GC_{3S}: the frequency of the nucleotides G+C at the 3rd of synonymous codons, GC_3: The G+C content at the 3rd of codons

Indices	GC	ENC	CAI	GC_{3S}	GC_3	Axis 1	Axis 2	Axis 3
ENC	0.099							
CAI	0.419**	-0.078						
GC3S	0.494**	0.531**	0.180					
GC3	0.545**	0.478**	0.250	0.922**				
Axis1	0.358*	-0.013	0.491**	0.085	0.098			
Axis2	-0.027	-0.148	0.245	0.299*	-0.131	-0.004		
Axis3	-0.100	0.012	0.038	0.030	0.077	-0.005	0.010	
Axis4	0.102	-0.221	0.035	-0.221	-0.108	-0.006	-0.008	0.007

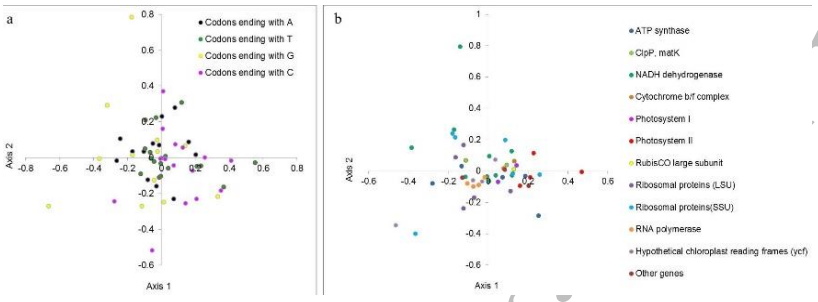

Fig. 4: Correspondence analysis of synonymous codon usage in yellowhorn chloroplast genome. The analysis is based on the RSCU values of 48 genes. a different base ended codons of Axis 2 versus Axis 1 are represented by different colors; b different gene types of Axis 2 versus Axis 1 are represented by different colors

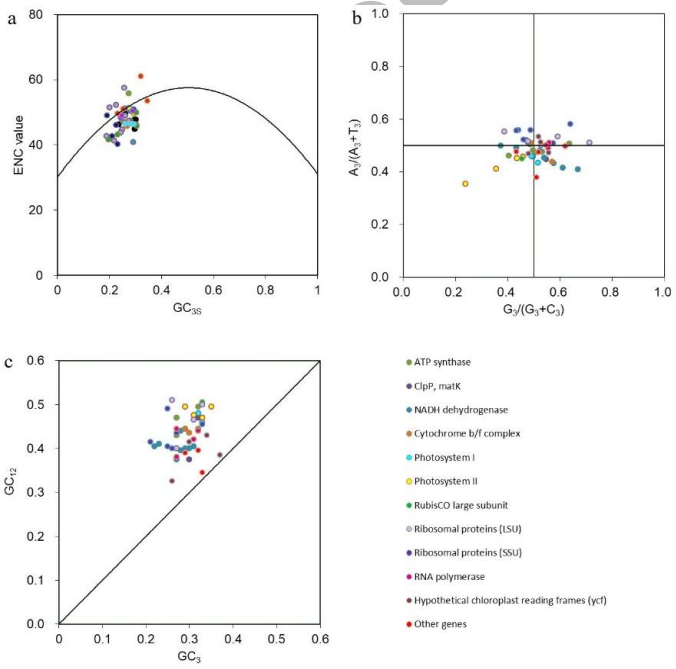

Fig. 5: Characteristics of evolutionary forces in yellowhorn. a ENC plot analysis of ENC values and GC_{3s} values; b PR2 plot analysis of the values $A_3/(A_3 + T_3)$ and $G_3/(G_3 + C_3)$, the curve shows the expected relationship between ENC values and GC_3 under random codon usage assumption; c Neutrality plot analysis of GC_{12} contents and GC_3 contents. The curve shows that GC_{12} is equal to GC_3

Foundation of China (Grant No. 31760242), and the Fundamental Research Funds for the Central Universities (Grant No. 31920190021).

Author contributions

XNG designed the research and performed the experiments.

XNG and YLW collected the data. XNG analyzed the data and wrote the manuscript. SMW revised the manuscript. All authors read and approved the manuscript.

References

Akashi H (1994). Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. *Genetics* 136:927–935

Bi QX, WB Guan (2014). Isolation and characterization of polymorphic genomic SSRs markers for the endangered tree *Xanthoceras sorbifolium* Bunge. *Conserv Genet Resour* 6:895–898

Chen X, X Cai, Q Chen, H Zhou (2011). Factors affecting synonymous codon usage bias in chloroplast genome of *Oncidium gower ramsey*. *Evol Bioinform* 7:271–278

Chen Y, Y Chen, C Shi, Z Huang, Y Zhang, S Li, Q Chen (2017). SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *GigaScience* 7:1-6

Comeron JM, M Aguadé (1998). An evaluation of measures of synonymous codon usage bias. *J Mol Evol* 47:268–274

Ding K, S Guo, WW Rong, Q Li, R Liu, H Xu, Y Yin, K Bi (2019). A new oleanane type pentacyclic triterpenoid saponin from the husks of *Xanthoceras sorbifolium* bunge and its neuroprotection on PC12 cells injury induced by Aβ25-35. *Nat Prod Res* 28:1–7

Dierckxsens N, P Mardulyn, G Smits (2016). NOVOPlasty: *de novo* assembly of organelle genomes from whole genome data. *Nucl Acids Res* 45; Article e18

Ghosh TC, SK Gupta, S Majumdar (2000). Studies on codon usage in *Entamoeba histolytica*. *Intl J Parasitol* 30:715–722

Goodwin S, J Gurtowski, S Ethe-Sayers, P Deshpande, MC Schatz, WR McCombie (2015). Oxford Nanopore sequencing, hybrid error correction, and *de novo* assembly of a eukaryotic genome. *Genome Res* 25:1750–1756

Guan DL, LB Ma, MS Khan, XX Zhang, SQ Xu, JY Xie (2018). Analysis of codon usage patterns in *Hirudinaria manillensis* reveals a preference for GC-ending codons caused by dominant selection constraints. *BMC Genomics* 19; Article 542

Hershberg R, DA Petrov (2008). Selection on codon bias. *Annu Rev Genet* 42:287–299

Ji XF, TY Chi, P Liu, LY Li, JK Xu, Q Xu, DL Meng (2017). The total triterpenoid saponins of *Xanthoceras sorbifolia* improve learning and memory impairments through against oxidative stress and synaptic damage. *Phytomedicine* 25:15–24

Katoh K, K Misawa, K Kuma, T Miyata (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucl Acids Res* 30:3059–3066

Kato T, T Kaneko, S Sato, Y Nakamura, S Tabata (2000). Complete structure of the chloroplast genome of a legume, *Lotus japonicus*. *DNA Res* 7:323–330

Kim KJ, HL Lee (2004). Complete chloroplast genome sequences from Korean Ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. *DNA Res* 11:247–261

Kumar S, G Stecher, M Li, C Knyaz, K Tamura (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol* 35:1547–1549

Li GL, ZL Pan, SC Gao, YY He, QY Xia, Y Jin, HP Yao (2019). Analysis of synonymous codon usage of chloroplast genome in *Porphyra umbilicalis*. *Genes Genomics* 41:1173–1181

Li X, YF Li, MY Zang, MZ Li, YM Fang (2018). Complete chloroplast genome sequence and phylogenetic analysis of *Quercus acutissima*. *Intl J Mol Sci* 19:2443–2459

Liang Q, HY Li, SK Li, FL Yuan, JF Sun, QC Duan, QY Li, R Zhang, YL Sang, N Wang, XW Hou, KQ Yang, JN Liu, L Yang (2019). The genome assembly and annotation of yellowhorn (*Xanthoceras sorbifolium* Bunge). *GigaScience* 8:1–15

Liu HM, R He, HY Zhang, YB Huang, ML Tian, JJ Zhang (2010). Analysis of SCU in *Zea mays*. *Mol Biol Rep* 37:677–684

Morton BR (1999). Strand asymmetry and codon usage bias in the chloroplast genome of *Euglena gracilis*. *Proc Natl Acad Sci USA* 96:5123–5128

Mukhopadhyay P, S Basak, TC Ghosh (2008). Differential selective constraints shaping codon usage pattern of housekeeping and tissue-specific homologous genes of rice and *Arabidopsis*. *DNA Res* 15:347–356

Perriere G, J Thioulouse (2002). Use and misuse of correspondence analysis in codon usage studies. *Nucl Acids Res* 30:4548–4555

Plotkin JB, G Kudla (2010). Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet* 12:32–42

Puigbò P, IG Bravo, S Garcia-Vallve (2008). CAIcal: A combined set of tools to assess codon usage adaptation. *Biol Direct* 3:38–45

Raubeson LA, R Peery, TW Chumley, C Dziubek, HM Fourcade, JL Boore, RK Jansen (2007). Comparative chloroplast genomics: analyses including new sequences from the angiosperms *Nuphar advena* and *Ranunculus macranthus*. *BMC Genomics* 8:174–200

Rice P, I Longden, A Bleasby (2000). EMBOSS: The European molecular biology open software suite. *Trends Genet* 16:276–277

Ruan CJ, R Yan, BX Wang, S Mopper, WK Guan, J Zhang (2017). The importance of yellow horn (*Xanthoceras sorbifolia*) for restoration of arid habitats and production of bioactive seed oils. *Ecol Eng* 99:504–512

Sato S, Y Nakamura, T Kaneko, E Asamizu, S Tabata (1999). Complete structure of the chloroplast genome of *Arabidopsis thaliana*. *DNA Res* 6:283–290

Sharp PM, E Cowe (1991). Synonymous codon usage in *Saccharomyces cerevisiae*. *Yeast* 7:657–678

Sharp PM, WH Li (1987). The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 15:1281–1295

Sueoka N (1999). 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* 238:53–58

Tao X, L Ma, Z Zhang, W Liu, Z Liu (2017). Characterization of the complete chloroplast genome of alfalfa *(Medicago sativa) (Leguminosae). Gene Rep* 6:67–73

Taylor FJ, D Coates (1989). The code within the codons. *Biosystems* 22:177–187

Wang HC, DA Hickey (2007). Rapid divergence of codon usage patterns within the rice genome. *BMC Evol Biol* 7:S6

Wang Q, L Yang, S Ranjitkar, JJ Wang, XR Wang, DX Zhang, ZY Wang, YZ Huang, YM Zhou, ZX Deng, LB Yi, XF Luan, YA El-Kassaby, WB Guan (2017). Distribution and in situ conservation of a relic Chinese oil woody species *Xanthoceras sorbifolium* (yellowhorn). *Can J For Res* 47:1450–1456

Wang Q, RB Zhu, JM Cheng, ZX Deng, WB Guan, YA El-Kassaby (2019a). Species association in *Xanthoceras sorbifolium* Bunge communities and selection for agroforestry establishment. *Agrofor Syst* 93:1531–1543

Wang X, Y Zheng, S Su, Y Ao (2019b). Discovery and Profiling of microRNAs at the Critical Period of Sex Differentiation in *Xanthoceras sorbifolium* Bunge. *Forests* 10:1141–1157

Wei L, J He, X Jia, Q Qi (2014). Analysis of codon usage bias of mitochondrial genome in *Bombyx mori* and its relation to evolution. *BMC Biol* 14; Article 262

Wold S, K Esbensen P Geladi (1987). Principal component analysis. *Chemometr Intell Lab* 2:37–52

Wright F (1990). The "effective number of codons" used in a gene. *Gene* 87:23–29

Zhou Q, Y Zheng (2015). Comparative *de novo* transcriptome analysis of fertilized ovules in *Xanthoceras sorbifolium* uncovered a pool of genes expressed specifically or preferentially in the selfed ovule that are potentially involved in late-acting self-incompatibility. *PLoS One* 10; Article e0140507