Journal of Agricultural Science; Vol. 9, No. 9; 2017
ISSN 1916-9752 E-ISSN 1916-9760
Published by Canadian Center of Science and Education

Codon Usage Bias of the Wheat Flower Development Gene *WAG-2* and Other AGAMOUS Group Genes

Wenhan Hu[1] & Shuhong Wei[1]

[1] Key Laboratory of Southwest China Wildlife Resources Conservation (Ministry of Education), China West Normal University, Nanchong, Sichuan, China

Correspondence: Shuhong Wei, College of Life Science, China West Normal University, No.1 Shi da Road, Nanchong 637009, Sichuan, China. Tel: 86-817-256-8315. E-mail: weishuhong453@sohu.com

Received: June 14, 2017 Accepted: July 22, 2017 Online Published: August 15, 2017

doi:10.5539/jas.v9n9p56 URL: https://doi.org/10.5539/jas.v9n9p56

The research is financed by the Dr. Start-up Foundation of China West Normal University (13E002).

Abstract

Analyzing codon usage bias of *WAG-2* gene in wheat three-pistil (TP) mutant may provide a basis for selecting the appropriate host expression systems to improve the expression of target genes. In the present study, we analyzed the codon bias of the complete coding sequence (CDS) of the *WAG-2* gene in TP using Codon W program, and compared the results with AGAMOUS (AG) group genes of other plant species. Results showed that the *WAG-2* gene in TP and other monocot AG group genes preferably used codons ending with G/C bases, but *Arabidopsis thaliana*, *Nicotiana tabacum*, and other dicot crops were biased toward the synonymous codons with A/T. The clustering results based on codon bias were consistent with those based on CDS of the AG group genes, indicating that the difference in codon preference of AG group genes sequences was closely associated with the genetic relationship of the species. The Euclidean distance coefficients of *WAG-2* with *A. thaliana* and *N. tabacum* were 9.255 and 5.730, respectively, indicating that *N. tabacum* may be more suitable for the expression of *WAG-2*. There were 37 codons showing distinct usage differences between *WAG-2* and genome of yeast, 23 between *WAG-2* and *Escherichia coli*. Therefore, the *E. coli* was the superior protein expression system. These results may improve our understanding of codon usage bias and functional studies of *WAG-2*.

Keywords: three-pistil mutation, *WAG-2*, AGAMOUS, codon usage bias, GC content, expression systems

1. Introduction

Common wheat (*Triticum aestivum* L.) line three-pistil (TP) mutation is a novel mutant of flower development, which was selected by Peng (2003) from the "tri-grain" wheat cultivar. The TP mutant stably carries three pistils in a floret and shows three normal stamens, making it a valuable material in wheat flower development.

The wheat AGAMOUS (AG) ortholog *WAG-2* is a C-class MADS-box gene. AG group genes play a central role in floral organ differentiation, formation, and development. Studies on genetic structure, expression and function of AG group genes have a potential application in seed plant breeding. Previous results using real-time PCR and in-situ hybridization have revealed that *WAG-2* might be associated with the development of pistil, ovule, and stamen homeotic transformation into pistil like structures (Mizumoto et al., 2009). However, the specific function of the *WAG-2* gene in floral organs, especially in the development of the pistil and ovule, remains unknown.

A total of 61 nucleotide codons are used to encode 20 amino acids and three codons to terminate translation. All amino acids are encoded by two to six synonymous codons with the exception of Met and Trp. Codon bias refers to the nonrandom usage of synonymous codons for encoding amino acids in organisms. After long term evolution, species form a set of specific codons to survive. The phenomenon of synonymous codon usage bias is widely observed in various species and genomes, even among different genes of the same genome (Ingvarsson, 2008; Liu, 2010). During the last few years, the synonymous codons usage in bacteria, yeast, and higher eukaryotes has been extensively analyzed. Brinkmann, the first to systematically analyze codon bias of monocots and dicot genes, has revealed that the codon usage between the chloroplast Gap^2 genes of maize and dicots is largely differences (Brinkmann et al., 1987). Study of 207 plant gene sequences confirms that codon usage in

nuclear genes differed between monocots and dicots (Murray et al., 1989). Compbell and Gowri (1990) compared codon usage in the genomes of cyanobacteria, green algae, and higher plants. They found proteins were encoded preferentially with codons ending in C or G in most cyanobacteria and the nuclei of green algae. A clear distinction between dicot and monocot codon usage is evident that dicot genes use preferred codons with a slight preference for codons ending A or U. Monocot codon usage is more restricted those ending in C or G. But two classes of genes could be recognized in monocots, one set of monocot genes uses codons similar to those in dicots, while the other genes are highly bias toward codons ending C or G. Grass genomes contain higher GC content compared with other angiosperm families (Šmarda & Bureš, 2012). In coding regions, GC content and G+C content in the 3rd codon position (GC3) in dicot exhibits a unimodal distribution, while it shows a bimodal distribution in grass genomes (Guo et al., 2007). Qi (2015) found that A- and T-ending codons are preferential in the plastid genomes of all 25 species from lower to higher plants. Recently, the force that influences the evolution of codon usage bias has drawn extensive attention of scholars. In bacteria, highly expressed genes appear to be under continuing strong selection, whereas selection is very weak in genes expressed at low levels (Sharp et al., 2010). Mutational bias as dominating force shapes the variation in the codon usage among the chloroplasts genes of pooid grass family (Sablok et al., 2011) and genes in *Citrus* and *Poncirus trifoliata* (Ahmad et al., 2013). To date, a large number of plant genes have been cloned and sequenced, which draw significant conclusions on their codon usage. Interest in higher plant codon usage has been heightened by the recognition that monocots differ from dicots in codons used to encode proteins with the same function in metabolism.

Codon usage bias within genes in a single species appears related to the level of expression of the protein encoded by that gene. Codon bias is most extreme in highly expressed proteins of *E.coli* and yeast. Hoekema (1987) reported that replacement of preferred codon by minor codons in the 5' end of the highly expressed yeast gene PGK1 resulted in a decreased level of both protein and mRNA. The bias codon choice in highly expressed genes enhances translation and is required for maintaining mRNA stability in yeast. The degree of codon bias may be a factor to consider when engineering high expression of heterologous genes in yeast and other system. If the exogenous genes have too many rare codons of expression systems and the preference differences between the exogenous gene and the expression system are significant, the transcription and translation levels of the exogenous gene in the host would be decreased.

In the present study, the codon usage bias of the *WAG-2* gene in TP was evaluated and compared with that of other AG group genes. Our findings provide a basis for selecting appropriate receptor plants and protein expression systems in *WAG-2* gene functional studies.

2. Method

2.1 Sequence Data

Four *WAG-2* gene transcripts, *WAG-2l*, *WAG-2m*, *WAG-2n* and *WAG-2o* (GenBank accession number: KT188782-KT188785) were cloned in our laboratory. The length of cDNA of the *WAG-2l*, *WAG-2m*, *WAG-2n* and *WAG-2o* genes are 1276 bp (coding for 273 amino acids), 1228 bp (274 amino acids), 1245 bp (275 amino acids), and 1242 bp (276 amino acids), respectively. Nineteen other AG group genes cDNA sequences were screened from the National Center for Biotechnology Information GenBank (http://www.ncbi.nlm.nih.gov/). The GenBank accession numbers for *WAG-2* and the 19 other AG group genes are listed in Table 1.

Table 1. The *WAG-2* gene and other AG group genes

Gene	Species	Accession No.	Gene	Species	Accession No.
WAG-1	*Triticum aestivum*	AB084577	*AG*	*Sorghum bicolor*	XM_002454940
WAG-2l	*T. aestivum*	KT188782	*PLE*	*Antirrhinum majus*	S53900
WAG-2m	*T. aestivum*	KT188783	*AG*	*Arabidopsis thaliana*	X53579
WAG-2m	*T. aestivum*	KT188784	*FBP7*	*Petunia hybrida*	X81651
WAG-20	*T. estivum*	KT188785	*FBP11*	*P. hybrida*	X81852
OsMADS3	*Oryza sativa*	L37528	*pMADS3*	*P. hybrida*	X72912
OsMADS58	*O. sativa*	FJ750942	*TAG1*	*Solanum lycopersicum*	L26295
ZAG1	*Zea mays*	NM-001111851	*BAG*	*Brassica napus*	M99415
ZMM2	*Z. mays*	L81162	*CAG1*	*Cucumis sativus*	AF022377
HvAG1	*Hordeum vulgare*	AF486648	*CAG2*	*C. sativus*	AF022378
HvAG2	*H. vulgare*	AF486649	NAG	*Nicotiana. tabacum*	L23925
MADS3	*Brachypodium distachyon*	XM_003565133			

2.2 Codon Usage Bias Analysis

Effective number of codon (ENC) can be used as a simple measure of codon bias in a gene, and is the best estimator of absolute synonymous codon usage bias. ENC value is between 20 (when only a single codon is used for each kind of amino acid, which means extreme preference) and 61 (when all available codons are used, signifying no bias). Low ENC value indicates a strong codon usage bias (Wright, 1990). The sequences with ENC values < 30 are highly expressed genes, whereas those with ENC values > 55 are poorly expressed genes (Biro, 2008). In addition, the GC and GC3 content of *WAG-2* and other AG group genes were calculated using Codon W 1.4 programs (http://codonw.sourceforge. net).

To investigate the characteristics of synonymous codon usage of different amino acid compositions, we calculated the relative synonymous codon usage (RSCU) values of 59 informative codons (excluding Met, Trp, and the three termination codons) in each CDS of AG group genes according to Codon W 1.4. The RSCU value was calculated by dividing the observed codon usage by the expected value when all codons for the same amino acid are used equally (Yang et al., 2010). If all synonyms for that amino acid are used equally, the RSCU values are close to 1.0, indicating a lack of bias. When the RSCU of a codon is more than 1.0, the codon has high-frequency usage (Sau et al., 2006).

Codon usage frequency is a measure of codon usage differences between species. Ratios between 0.5 and 2.0 show that the biases of the two codons are relatively close. Ratios equal to or less than 0.5 and equal to or greater than 2.0 indicate that the codon usages are different. Choosing a host is important in transgenic research, and choosing the appropriate codons is one of the most vital factors that affect expression in hosts. If foreign genes contain numerous rare codons that are incompatible with the expression system of the host, the result is extremely low expression quantity or termination of translation, especially when rare codons are distributed continuously.

2.3 Statistical Analysis

The comparison between RSCU value and 1.0 was performed with one-sample t-test. The differences of ENC, GC, GC3 between monocot and dicot plants were calculated using Kruskal-Wallis test.

2.4 Clustering Analysis

First, we used the Euclidean square distance to conduct clustering analysis based on RSCU values of the *WAG-2* and other AG group genes after data standardization. The formula used to calculate the Euclidean distance coefficient (D*ab*) of codon usage bias between two genes a and b is as follows:

$$Dab = \sqrt{\sum_{i=1}^{59} (RSCUai - RSCUbi)^2} \quad (1)$$

3. Results

3.1 Synonymous Codon Usage of WAG-2 Genes

The ENC, GC, and GC3 content of AG group genes were shown in Table 2. The synonymous codon usage of four *WAG-2* genes in TP was remarkably close. The ENC values of *WAG-2l*, *WAG-2m*, *WAG-2n*, and *WAG-2o* were significantly lower than 55 (41.54, 42.22, 41.71, and 41.78, respectively), suggesting that *WAG-2* genes in TP had moderate preference in codon usage and the expression levels were general. The GC content of *WAG-2l*, *WAG-2m*, *WAG-2n*, and *WAG-2o* were 0.559, 0.554, 0.559, and 0.562, respectively, with a mean value of 0.558. The GC3 content were significantly greater than 0.5 (average of 0.764), indicating that *WAG-2* genes preferably used G-ended or C-ended codons.

The RSCU values of 59 codons in four *WAG-2* genes were analyzed. Given that the codon usage of the four *WAG-2* genes was highly consistent, only the average RSCU value was listed in Table 3. In the *WAG-2l*, *WAG-2m*, *WAG-2n*, and *WAG-2o* genes, there were 23, 23, 24, and 24 codons, respectively, whose RSCU values were significantly higher than 1.0 ($p < 0.01$). Therefore they are the optimal codons of the *WAG-2* genes. In addition to the GAT and CAT, the rest of the 21 codons preferably used G- or C-based endings. The faction and frequency values of these codons were higher, which verified that the *WAG-2* genes preferred the G or C ends.

3.2 Composition with Other AG Group Genes on Codon Usage Bias

Table 2 showed the codon usage bias of AG group genes in monocots and dicots. In monocots, the ENC values of *WAG-1* (*T. aestivum*), *OsMADS58* (*O. sativa*) were slightly lower than 55 (53.36 and 53.80, respectively), suggesting no obvious preference in codon usage. The ENC values in other monocots varied from 42.41 to 49.79, with a mean value of 46.21, indicating that these AG group genes were moderately biased. The ENC values of AG group genes in all monocot species were lower than 55, suggesting that the expression levels of AG group

genes in these species were general. In dicots, only one AG group gene *CAG1* (*C. sativus*) had lower ENC value (47.35) than the other species. The ENC values in other species were relatively higher than 50 with a mean of 56.46, suggesting no obvious preference in codon usage and that the expression levels of AG group genes in these species were poor.

Interestingly, the contents of GC3 (average of 0.7165) in monocots were significantly higher than those in dicots (0.4649). GC contents in monocots (> 0.5) were different from those in dicots (< 0.5). In monocots, the average percentage of G+C at the third position of the codon was the highest, reaching a maximum of 94.30%. However, the contents of G+C and A+T in dicots were 69.46% and 62.25%, respectively.

Our results from the Kruskal-Wallis test showed that the ENC value were significantly different between monocot and dicot plants ($p < 0.05$), exception for *WAG-1* and *OsMADS58*. The GC and GC3 content in monocot was significantly higher compared with dicot ($p < 0.05$). These results showed a significant difference in G or C preference and in ENC, GC, and GC3 contents in both monocot and dicot plants. AG group genes preferred C-ended or G-ended codons at the synonymous positions in monocots, whereas ending with A or T in dicots.

Table 2. The ENC values and contents of GC for *WAG-2* genes and other AG group genes

Gene	ENC	GC	T3	C3	A3	G3	GC3
WAG1	53.36	0.514	0.1359	0.4674	0.2132	0.4892	0.725
WAG-2l	41.54	0.558	0.1800	0.5000	0.1154	0.5051	0.780
WAG-2m	42.22	0.552	0.1894	0.4862	0.1202	0.5051	0.767
WAG-2m	41.74	0.558	0.1850	0.4900	0.1148	0.5176	0.779
WAG-2o	41.78	0.561	0.1872	0.4828	0.1132	0.5149	0.776
OsMADS3	44.96	0.516	0.2241	0.4598	0.1854	0.4524	0.684
OsMADS58	53.80	0.509	0.2124	0.4495	0.2488	0.3673	0.617
ZAG1	49.79	0.542	0.1991	0.4769	0.1689	0.4524	0.710
ZMM2	43.60	0.562	0.1458	0.5260	0.1310	0.4948	0.780
HvAG1	46.90	0.537	0.1941	0.5176	0.1356	0.4675	0.746
HvAG2	46.89	0.517	0.1726	0.4762	0.2191	0.4491	0.695
MADS3	45.74	0.560	0.1852	0.5185	0.1294	0.4639	0.755
AG	46.12	0.554	0.1592	0.5025	0.1739	0.4462	0.734
PLE	56.66	0.452	0.3801	0.2954	0.2787	0.3663	0.493
AG	58.23	0.432	0.3591	0.3091	0.3591	0.2687	0.436
FTP7	56.11	0.428	0.3585	0.2186	0.3396	0.4049	0.468
FTP11	51.31	0.427	0.3625	0.2313	0.3631	0.3675	0.443
PMADS3	57.88	0.426	0.3867	0.2928	0.3441	0.2965	0.437
TAG1	54.56	0.426	0.3520	0.2905	0.3817	0.3103	0.442
BAG1	61.00	0.443	0.3333	0.3228	0.3731	0.2697	0.447
CAG1	47.35	0.404	0.4268	0.2561	0.3652	0.2515	0.381
CAG2	51.52	0.395	0.3737	0.2273	0.4564	0.2278	0.345
NAG	53.08	0.423	0.3778	0.2833	0.3542	0.3146	0.440

Note. ENC, effective number of codon; GC, G+C content; GC3, G+C content in the 3rd codon position; A3, A content in the 3rd codon position; C3, C content in the 3rd codon position; G3, G content in the 3rd codon position; T3, T content in the 3rd codon position.

3.3 Comparison with Genomes of E. coli and Yeast on Codon Usage Frequency

The differences in *WAG-2* codon usage in various hosts affect expression levels. Thus, codon preference must be considered when genes are expressed in heterologous hosts. The usage frequencies of 64 codons in *WAG-2* were compared with those in *E. coli* and yeast (Table 4). We found that the number of codons with ratios > 2.0 and/or < 0.5 were 23 in *E. coli* and 37 in yeast. This result suggested that the *E. coli* expression system may be superior to the yeast expression system for *WAG-2*.

3.4 Clustering Analysis

Two types of common model plants, *A. thaliana* and *N. tabacum*, are widely used in the study of plant gene expression and function. To discover whether *WAG-2* can be expressed efficiently in the two model plants, we

conducted clustering analysis based on codon bias using the squared Euclidean distance method (Figure 1). The Euclidean distance coefficients of the *WAG-2* gene with *A. thaliana* and *N. tabacum* were 9.255 and 5.730, respectively, which indicated that the codon usage bias between the *WAG-2* and *NAG* genes was more similar (Appendix A). Thus the *N. tabacum* may be more suitable for the heterogeneous expression system of the *WAG-2* gene. In addition, the resultant cluster was clearly classified in monocot and dicot clades according to their codon usage bias. *T. aestivum*, *H. vulgare*, *B. distachyon*, *O.sativa*, *Z. mays*, and *S. bicolor* were included in the monocot clade. The dicot clade consisted of *A. thaliana*, *A. majus*, and *P. hybrid*. Monocot genes were subdivided into three clades. *WAG-2*, *HvAG1*, and *MADS3* were clustered in one subclade. *ZMM2*, *AG* (*S. bicolor*), and *OsMADS3* were attributed to another subclade. *OsMADS58*, *ZAG1*, *WAG-1*, and *HvAG2* were included in the third subclade. The phylogenetic tree based on CDS indicated that the AG group was also classified into monocot and dicot clades (Figure 2). Both phylogenetic trees were highly similar to each other and differed only in the positions of *CAG1* and *CAG2* in the dicot clade.

Table 3. The RSCU value of *WAG-2* gene

AA	Codon	Fraction	Frequency	RSCU	AA	Codon	Fraction	Frequency	RSCU
A(Ala)	GCT	0.250	18.962	1.00	P(Pro)	CCT	0.000	0.000	0.00
A	GCC	0.418	31.756	1.67	P	CCC	0.000	0.000	0.00
A	GCA	0.048	3.630	0.19	P	CCA	0.175	3.630	0.70
A	GCG	0.285	21.782	1.14	P	CCG	0.825	17.237	3.30
C(Cys)	TGT	0.000	0.000	0.00	Q(Gln)	CAA	0.078	7.250	0.05
C	TGC	1.000	10.890	2.00	Q	CAG	0.922	86.200	1.85
D(Asp)	GAT	0.675	24.493	1.35	R(Arg)	CGT	0.000	0.000	0.00
D	GAC	0.325	11.805	0.65	R	CGC	0.281	20.870	1.68
E(Glu)	GAA	0.263	19.173	0.52	R	CGA	0.013	0.909	0.07
E	GAG	0.737	53.538	1.48	R	CGG	0.109	8.162	0.66
F(Phe)	TTT	0.000	0.000	0.00	R	AGA	0.146	10.889	0.88
F	TTC	0.700	14.519	1.90	R	AGG	0.451	33.571	2.71
G(Gly)	GGT	0.154	7.259	0.62	S(Ser)	TCT	0.099	10.889	0.60
G	GGC	0.500	23.601	2.00	S	TCC	0.265	29.039	1.59
G	GGA	0.115	5.438	0.46	S	TCA	0.075	8.172	0.42
G	GGG	0.231	10.879	0.92	S	TCG	0.132	14.519	0.79
H(His)	CAT	0.729	9.977	1.45	S	AGT	0.124	13.613	0.74
H	CAC	0.270	3.630	0.54	S	AGC	0.306	33.591	1.83
I(Ile)	ATT	0.200	7.259	0.60	T(Thr)	ACT	0.068	3.636	0.37
I	ATC	0.600	21.779	1.80	T	ACC	0.677	26.321	2.61
I	ATA	0.200	7.259	0.60	T	ACA	0.093	3.630	0.47
K(Lys)	AAA	0.200	10.889	0.40	T	ACG	0.161	6.347	0.65
K	AAG	0.800	43.558	1.60	V(Val)	GTT	0.138	7.260	0.55
L(Leu)	TTA	0.000	0.000	0.00	V	GTC	0.345	18.149	1.38
L	TTG	0.081	7.259	0.55	V	GTA	0.000	0.000	0.00
L	CTT	0.088	7.260	0.55	V	GTG	0.516	27.230	2.06
L	CTC	0.264	21.779	1.59	W(Trp)	TGG	0.000	0.000	0.00
L	CTA	0.088	7.260	0.53	Y(Tyr)	TAT	0.215	9.980	0.43
L	CTG	0.475	36.128	2.83	Y	TAC	0.785	36.299	1.57
M(Met)	ATG	1.000	46.285	1.00	*	TAA	0.000	0.000	*
N(Asn)	AAT	0.227	15.431	0.45	*	TAG	1.000	3.650	*
N	AAC	0.773	52.632	1.55					

Note. *Termination codon. The data with underline mean that RSCU > 1.0.

Table 4. Codon frequency of *WAG-2* genes in TP, *E. coli* genome and yeast genome (f/‰)

AA	Codon	*WAG-2*	*E.coli* genome	yeast genome	*WAG-2*/*E.coli* genome	*WAG-2*/yeast genome
A(Ala)	GCT	18.962	15.6	21.2	1.22	0.89
A	GCC	31.756	25.1	12.6	1.27	2.52
A	GCA	3.630	20.6	16.2	0.17	0.22
A	GCG	21.782	31.7	6.2	0.69	3.51
C(Cys)	TGT	0.000	5.5	8.1	0.00	0.00
C	TGC	10.890	6.9	4.8	1.58	2.27
D(Asp)	GAT	24.493	32.1	37.8	0.76	0.65
D	GAC	11.805	18.6	20.2	0.63	0.58
E(Glu)	GAA	19.173	38.2	45.6	0.50	0.42
E	GAG	53.538	17.7	19.2	3.02	2.79
F(Phe)	TTT	0.000	23.2	26.1	0.00	0.00
F	TTC	14.519	16.9	18.4	0.86	0.79
G(Gly)	GGT	7.259	24.4	23.9	0.30	0.30
G	GGC	23.601	27.9	9.8	0.85	2.41
G	GGA	5.438	9	10.9	0.60	0.50
G	GGG	10.879	11.3	6.0	0.96	1.81
H(His)	CAT	9.977	13.6	13.6	0.73	0.73
H	CAC	3.630	9.8	7.8	0.37	0.47
I(Ile)	ATT	7.259	29.8	30.1	0.24	0.24
I	ATC	21.779	24.2	17.2	0.73	1.27
I	ATA	7.259	5.4	17.8	1.34	0.41
K(Lys)	AAA	10.889	33.2	41.9	0.33	0.26
K	AAG	43.558	10.7	30.8	4.07	1.41
L(Leu)	TTA	0.000	13.9	26.2	0.00	0.00
L	TTG	7.259	14.0	27.2	0.52	0.27
L	CTT	7.260	11.7	12.3	0.62	0.59
L	CTC	21.779	11.0	5.4	1.98	4.03
L	CTA	7.260	4.0	13.4	1.82	0.54
L	CTG	36.128	50.9	10.5	0.71	3.44
M(Met)	ATG	46.285	27.9	20.9	1.66	2.21
N(Asn)	AAT	15. 431	18.8	35.7	0.82	0.43
N	AAC	52.632	21.4	24.8	2.46	2.12
P(Pro)	CCT	0.000	7.3	13.5	0.00	0.00
P	CCC	0.000	5.8	6.8	0.00	0.00
P	CCA	3.630	8.5	18.3	0.43	0.20
P	CCG	17.237	21.8	5.3	0.79	3.25
Q(Gln)	CAA	7.250	15.0	27.3	0.48	0.27
Q	CAG	86.200	29.5	12.1	2.92	7.12
R(Arg)	CGT	0.000	20.3	6.4	0.00	0.00
R	CGC	20.870	21.0	2.6	0.99	8.03
R	CGA	0.909	3.9	3.0	0.00	0.00
R	CGG	8.162	6.3	1.7	1.30	4.80
R	AGA	10.889	2.9	21.3	3.75	0.51
R	AGG	33.571	1.9	9.2	17.6	3.65
S(Ser)	TCT	10.889	8.7	23.5	1.25	0.46
S	TCC	29.039	8.9	14.2	3.26	2.05
S	TCA	8.172	7.8	18.7	1.05	0.44
S	TCG	14.519	8.7	8.6	1.67	1.69
S	AGT	13.613	9.5	14.2	1.43	0.67

S	AGC	33.591	16.0	9.8	2.10	3.43
T(Thr)	ACT	3.636	9.1	20.3	0.40	0.18
T	ACC	26.321	22.8	12.7	1.15	2.07
T	ACA	3.630	8.2	17.8	0.44	0.20
T	ACG	6.347	14.8	8.0	0.43	0.80
V(Val)	GTT	7.260	18.5	22.1	0.39	0.33
V	GTC	18.149	15.1	11.8	1.20	1.54
V	GTA	0.000	11.1	11.8	0.00	0.00
V	GTG	27.230	25.5	10.8	1.07	2.52
W(Trp)	TGG	0.000	15.2	10.3	0.00	0.00
Y(Tyr)	TAT	9.980	16.5	18.8	0.60	0.53
Y	TAC	36.299	12.1	14.8	3.00	2.45
*	TAA	0.000	2.0	1.0	0.00	0.00
*	TAG	3.650	0.3	0.5	12.17	7.30
*	TGA	0.000	1.1	0.7	0.00	0.00

Note. *Termination codon. The data with underline refer obvious differences of value (≤ 0.5, > 2.0) about the codon bias.

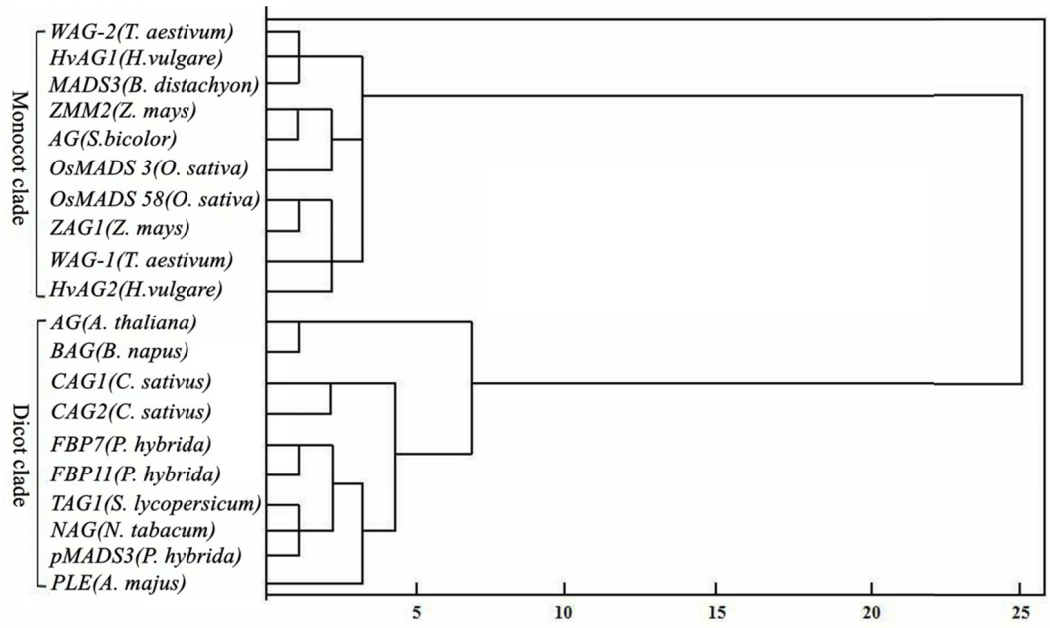

Figure 1. Clustering dendrogram based on RSCU of *WAG-2* gene and other AG group genes

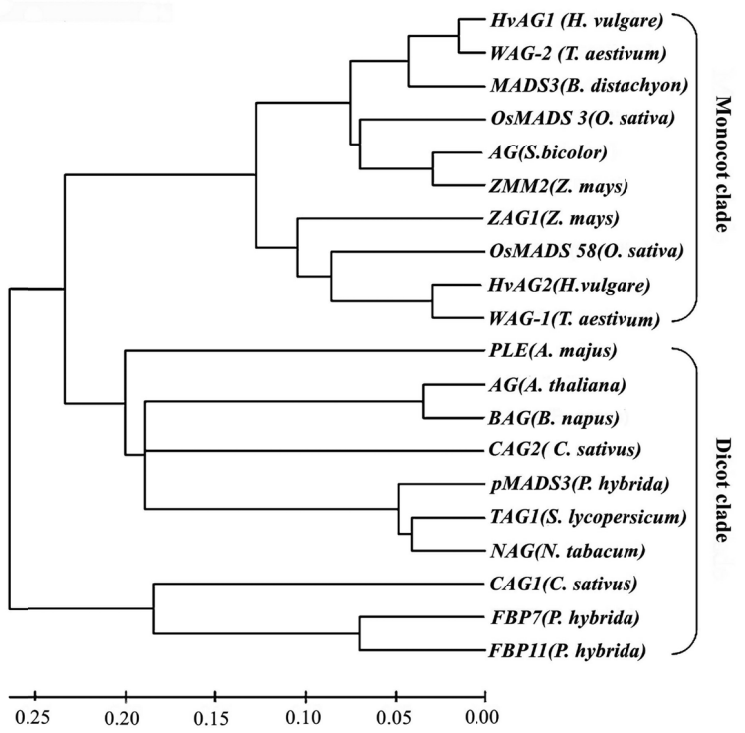

Figure 2. Clustering based on CDS of *WAG-2* and other AG group genes

4. Discussion

Codon usage is different for each gene or genome mainly occurring in choices between codons ending in base C or G verse those ending in A or U. GC3 is a useful proxy to understand the forces affecting genomic GC content. GC content is a characteristic feature of genome organization, varies between genes within a genome, and between genomes of different organisms. Earlier reports revealed that the codon profile in monocots was more strikingly bias in its preference for G/C, while dicots favored A/T in this position (Murray et al., 1989). Grass genomes contained higher GC content compared with other angiosperm families (Šmarda & Bureš, 2012). It was known that GC and GC3 content in dicot was unimodal, while it showed a bimodal distribution in grass genomes (Guo et al., 2007). Bimodal distribution of GC content in grass genomes has resulted from an increase in GC content from 5' to 3' among genes within genomes (Adrienne et al., 2015). In addition, genes with high GC3 content provide more targets for methylation, exhibit more variable expression, more frequently possess upstream TATA boxes and are predominant in certain classes of genes (Tatarinova et al., 2010). The forces that influence the GC content variation have long been considered to be nature selection (Sharp et al., 2010), mutational biases (Sablok et al., 2013; Sablok et al., 2011) and GC-biased gene conversion (gBGC) (Lassalle et al., 2015). Previous study showed that plant MADS-box type II genes, including AG group genes were highly conserved transcription factors. Type II genes have experienced a lower rate of birth-and-death evolution partly due to a stronger purifying selection (Nam et al., 2004). And *WAG-2* in wheat and its relatives belonged to a conserved gene affected by negative selection (Wei et al., 2011). Given these we conclude that mutation pressure should be the primary influence on high GC content of AG group genes. The mutation pressure can affect the base composition of a sequence in certain directions without natural selection pressure. This effect is reflected in the GC content and the third base of synonymous codons (Chen et al., 2004). If the mutation pressure of AT is higher than GC, then the third codon becomes either G or C (Novembre, 2002). In the present study, wheat *WAG-2* and other monocot AG group genes had high GC content (Table 3) and preferred to use codons with C or G at the synonymous position, whereas dicot AG group genes preferred the codons ending with A or T. Thus, we conclude that the mutation pressure of GC probably due to the stronger stacking interaction between GC base pair is lower than that of AT in *WAG-2* and other AG group genes of most monocots during evolution. In addition, according to gBGC, elevated GC content is caused by preferential incorporation of GC alleles during double-strand break repair by homologous recombination. In human and other genome, there has been some strong evidence to support the gBGC hypothesis (Ratnakuma, 2010). To our knowledge, not much is known

about the correlation between homologous recombination and GC content of AG group genes in monocot, which needs further discussion.

Genes with different functions also possessed distinct codon usage patterns in plants (Rota-Stabelli et al., 2013). High ENC values and low ENC genes belonging to the same gene family share different functions (Liu et al., 2015). In the present study, the ENC value of the *WAG-2* gene was lower (with a mean of 41.78) than that of the *WAG-1* gene (53.36). Previous studies indicated that AG group C-function gene was grouped into two AG orthologs, where wheat *WAG-1* and *WAG-2* belonged to AG1 and AG2 orthologs, respectively. Minimal *WAG-2* genes were found in transformed and developing stamens at the floral organ development stage, but were abundant in the marginal region of the developing ovule and in the central region of pistils (Mizumoto et al., 2009). Nevertheless, the *WAG-1* gene was associated with the development of pistil and stamen, and with pistillody caused by nuclear-cytoplasm interactions in alloplasmic wheat (Meguro et al., 2013). Wheat AG orthologs *WAG-1* and *WAG-2* exhibited functional differentiation during floral organ development. This phenomenon showed that the codon usage bias pattern of wheat *WAG-1* and *WAG-2* genes was somehow linked to gene function, which requires further investigation.

Understanding the codon usage bias can show the codon usage pattern of species, and provide evidence about the evolution of organisms. Higher plants are like other organisms in that each species has a unique codon bias with plants of the same taxonomic class maintaining a similar codon usage pattern (Campbell & Gowri, 1990). Species with near genetic relationship share the near codon usage frequency and preference. In the present study, codon usage of AG group genes in 10 monocot and 10 dicots were analyzed. It was found that the relationships of species were more closely, and the codon usage patterns of AG group genes were more similar. The genetic relationships between monocots were close, and their AG group genes on codon usage were also similar, so did dicots. The clustering results based on codon usage bias were consistent with those based on the CDS of the AG group genes (Figures 1 and 2). These results indicated that the difference in codon preference of AG group genes was closely associated with the genetic relationship of the species. So the analysis of codons usage bias was an important and supplementary method to phylogenetic research, and was used to the investigation of the evolutionary relationships of species.

In dicots, *A. thaliana* and *N. tabacum* were two types of general receptor plant in the study of gene expression and function. Efficient expression of exogenous genes in *E. coli* or yeast will lay a foundation for the identification of gene function. The codons usage bias in highly expressed genes enhances translation and is required for maintaining mRNA stability in yeast. The degree of codon bias may be a factor to consider when engineering high expression of heterologous genes in yeast and other system. Species within the same taxonomic class exhibit a similar codon usage pattern. And species with near genetic relationship share the near codon usage frequency and preference. In the present study, the codon usage bias between the *WAG-2* and *N. Tabacum NAG* genes was more similar. Consequently, we concluded that *N. tabacum* was the superior heterologous expression systems, which required further study. Compared with yeast and *E.coli* genome, *WAG-2* gene showed difference of 37 and 23, respectively, indicating *E.coli* was the superior protein expression system. If the *WAG-2* gene showed a high expression level in yeast, some modifications on different partial codons would be required.

5. Conclusion

In summary, the codon usage patterns and phylogenetic information provided in this study may help in determining the appropriate expression system of exogenous and in investigating the function of *WAG-2* genes in TP.

References

Adrienne, R., Sylvain, G., Pierre, M., Laurana, S. G., Christine, D., & Johann, J. (2015). Introns structure patterns of variation in nucleotide composition in *Arabidopsis thaliana* and rice protein-coding genes. *Genome Biology & Evolution, 7*(10), 2913-2928. https://doi.org/10.1101/010819

Ahmad, T., Sablok, G., Tatarinova, T. V., Xu, Q., Deng, X. X., & Guo, W. W. (2013). Evaluation of codon biology in citrus and Poncirus trifoliata based on genomic features and frame corrected expressed sequence tags. *DNA Research an International Journal for Rapid Publication of Reports on Genes & Genomes, 20*(2), 135-150. https://doi.org/10.1093/dnares/dss039

Biro, J. C. (2008). Does codon bias have an evolutionary origin? *Theoretical Biology and Medical Modelling, 5*(1), 3452. https://doi.org/10.1186/1742-4682-5-16

Brinkmann, H., Martinez, P., Quigley, F., Martin, W., & Cerff, R. (1987). Endosymbiotic origin and codon bias of the nuclear gene for chloroplast glyceradehyde-3-phosphate dehydrogenase from maize. *Journal of Molecular Evolution, 26*(4), 320-328. https://doi.org/10.1007/BF02101150

Campbell, W. H., & Gowri, G. (1990). Codon Usage in Higher Plants, Green Algae, and Cyanobacteria. *Plant Physiology, 92*(92), 1-11. https://doi.org/10.1104/pp.92.1.1

Chen, S. L., Lee, W., Hottes, A. K., Shapiro, L., & Mcadams, H. H. (2004). Codon usage between genomes is constrained by genome-wide mutational processes. *Proceedings of the National Academy of Sciences of the USA, 101*(10), 3480-3485. https://doi.org/10.1073/pnas.0307827100

Guo, X. Y., Bao, J. D., & Fan, L. J. (2007). Evidence of selectively driven codon usage in rice: Implications for GC content evolution of Gramineae genes. *Febs Letter, 581*(5), 1015-1021. https://doi.org/10.1016/j.febslet.2007.01.088

Hoekma, A, Kastelein, R. A., Vasser, M., & de Boer, H. A. (1987). Codon replacement in the PGK1 gene of Saccharomyces cerevisiae: Experimental approach to study the role of biased codon usage in gene expression. *Molecular and Cellular Biology, 7*(8), 2914-2924. https://doi.org/10.1128/MCB.7.8.2914

Ingvarsson, P. K. (2008). Molecular evolution of synonymous codon usage in *Populus*. *BMC Evolutionary Biology, 8*(1), 1-13. https://doi.org/10.1186/1471-2148-8-307

Lassalle, F., Périan, S, Bataillon, T., Nesme, X., Duret, L., & Daubin, V. (2015). GC-content evolution in bacterial genomes: The biased gene conversion hypothesis expands. *PloS Genetic, 11*(2), e1004941. https://doi.org/10.1371.journal.pgen.e1004941

Li, W. H. (1987). Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. *Journal of Molecular Evolution, 24*(4), 337-345. https://doi.org/10.1007/BF02134132

Liu, Q. P. (2010). Mutational bias and translational selection shaping the codon usage pattern of tissue-specific genes in rice. *PloS One, 7*(7), e48295. https://doi.org/10.1371/journal.pone.0048295

Liu, Q. P., Hu, H. C., Wang, H. (2015). Mutational bias is the driving force for shaping the synonymous codon usage pattern of alternatively spliced genes in rice (*Oryza sativa* L.). *Molecular Genetic and Genomics, 290*(2), 649-660. https://doi.org/10.1007/s00438-014-0951-0

Meguro, A., Takumi, S., Ogihara, Y., & Murai, K. (2013). *WAG*, a wheat AGAMOUS homolog, is associated with development of pistil-like stamens in alloplasmic wheats. S*ex Plant Reproduce, 15*(5), 221-230. https://doi.org/10.1007/s00497-002-0158-0

Mizumoto, K., Hatano, H., Hirabayashi, C., Murai, K., & Takumi, S. (2009). Altered expression of wheat AINTEGUMENTA homolog, *WANT-1*, in pistil and pistil-like transformed stamen of an alloplasmic line with *Aegilops crassa* cytoplasm. *Development. Genes and Evolution, 219*(4), 75-187. https://doi.org/10.1007/s00427-009-0275-y

Nam, J., Kim, J., Lee, S., An, G., Ma, H., & Nei, M. (2004). Type I MADS-box genes have experienced faster birth-and-death evolution than type II MADS-box genes in angiosperms. *Proceedings of the National Academy of Sciences, USA, 101*(7), 1910-1915. https://doi.org/10.1073pnas.0308430100

Novembre, J. A. (2002). Accounting for background nucleotide composition when measuring codon usage bias. *Molecular Biology & Evolution, 19*(8), 1390-1394. https://doi.org/10.1093/oxfordjournals.molbev.a004201

Pan, L. L., Wang, Y., Hu, J. H., Ding, Z. T., & Li, C. (2013). Analysis of codon use features of stearoyl-acyl carrier protein desaturase gene in *Camellia sinensis*. *Journal of Theoretical Biology, 334*(19), 80-86. https://doi.org/10.1016/j.jtbi.2013.06.006

Peng, Z. S. (2003). A new mutation in wheat producing three pistils in a floret. *Journal of Agronomy & Crop Science, 189*(4), 270-272. https://doi.org/10.1046/j.1439-037X.2003.00040.x

Qi, Y. Y., Xu, W. J., Xing, T., Zhao, M. M., Li, N. N., Yan, L., … Wang, M. M. (2015). Synonymous codon usage bias in the plastid genome is unrelated to gene structure and shows evolutionary heterogeneity. *Evolutionary Bioinformatics Online, 11*(11), 65-77. https://doi.org/10.4137/EBO.s22566

Ratnakumar, A., Mousset, S., Glémin, S., Berglund, J., Galtier, N., Duret, L., & Webster, M. T. (2010). Detecting positive selection within genomes: the problem of biased gene conversion. *Philosophical Transactions of the Royal Society B Biological Sciences, 365*(1652), 2571-2580. https://doi.org/10.1098/rstb.2010.0007

Rota-Stabelli, O., Lartillot, N., Philippe, H., & Pisani, D. (2013). Serine codon-usage bias in deep phylogenomics: Pancrustacean relationships as a case study. *Systematic Biology, 62*(1), 268-269. https://doi.org/10.1093/sysbio/sys077

Sablok, G., Nayak, K. C., Vazquez, F., & Tatarinova, T. V. (2011). Synonymous codon usage, gc(3), and evolutionary patterns across plastomes of three pooid model species: Emerging grass genome models for monocots. *Molecular Biotechnology, 49*(2), 116-128. https://doi.org/10.1007/s12033-011-9383-9

Sau, K., Gupta, S. K., Sau, S., Mandal, S. C., & Ghosh, T. C. (2006). Factors influencing synonymous codon and amino acid usage biases in Mimivirus. *Biosystems, 85*(2), 107-113. https://doi.org/10.1016/j.biosystems.2005.12.004

Sharp, P. M., Emery, L. R., & Zeng, K. (2010). Forces that influence the evolution of codon bias. *Philosophical Transactions of the Royal Society of London, 365*(1544), 1203-1212. https://doi.org/10.1098/rstb.2009.0305

Šmarda, P., & Bureš, P. (2012). The Variation of base composition in plant genomes. In J. F. Wendel, J. Greilhuber, J. Dolezel, & I. J. Leitch (Eds.), *Plant genome diversity* (Vol. 1, pp. 209-235). Springer Press. https://doi.org/10.1007/978-3-7091-1130-7_14

Tatarinova, T. V., Alexandrov, N. N., Bouck, J. B., & Feldmann, K. A. (2010). GC3 biology in corn, rice, sorghum and other grasses. *BMC Genomics, 11*(1), 144. https://doi.org/10.1186/1471-2164-11-308

Wei, S. H., Peng, Z. S., Zhou, Y. H., Yang, Z. J., Wu, K., & Ouyang, Z. M. (2011). Nucleotide diversity and molecular evolution of the *WAG-2* gene in common wheat (*Triticum aestivum* L.) and its relatives. *Genetics & Molecular Biology, 34*(3), 606-615. https://doi.org/10.1590/S1415-47572011000400013

Wright, F. (1990). The 'effective number of codons' used in a gene. *Gene, 87*(1), 23-29. https://doi.org/10.1016/0378-1119(90)90491-9

Yang, J., Zhu, T. Y., Jiang, Z. X., Chen, C., & Wang, Y. L. (2010). Codon usage biases in Alzheimer's disease and other neurodegenerative diseases. *Protein Peptide. Letter., 17*(5), 630-645. https://doi.org/10.2174/092986610791112666

Appendix

Appendix A. Coefficient of absolute squared Euclidean distance of codon usage bias among the sample of AG group genes

Species	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0.000																			
2	3.408	0.000																		
3	5.660	2.609	0.000																	
4	4.100	3.212	3.158	0.000																
5	4.346	2.688	2.386	2.140	0.000															
6	2.671	2.325	4.295	3.560	3.251	0.000														
7	3.333	1.168	3.096	3.462	3.350	2.727	0.000													
8	3.409	2.991	4.008	3.358	2.352	2.720	3.663	0.000												
9	3.138	1.508	2.321	3.290	2.758	2.017	1.288	2.182	0.000											
10	3.270	2.696	3.220	2.344	2.251	1.923	2.856	3.878	2.711	0.000										
11	7.626	7.462	6.252	4.688	4.820	7.292	7.381	4.729	7.123	7.156	0.000									
12	9.663	9.255	7.559	6.406	6.710	9.955	8.958	7.728	8.518	8.347	4.837	0.000								
13	7.688	6.620	5.372	3.949	4.640	8.639	7.183	5.356	6.691	7.713	4.390	4.088	0.000							
14	8.445	6.548	5.163	2.294	4.319	8.642	7.310	5.742	6.975	6.493	4.031	5.213	1.156	0.000						
15	8.150	8.044	7.736	3.808	4.850	8.308	7.790	6.210	7.725	7.028	3.590	5.275	3.740	3.181	0.000					
16	7.451	6.706	5.393	3.090	4.857	8.049	7.271	5.636	6.535	7.098	3.723	4.984	2.309	2.784	2.425	0.000				
17	8.114	10.719	8.944	6.874	7.926	10.350	9.345	8.370	8.950	9.186	5.068	2.486	4.940	6.290	4.928	5.134	0.000			
18	9.719	9.000	7.505	5.663	5.618	9.490	9.206	5.596	8.739	9.228	4.321	5.748	2.723	3.658	4.419	4.232	6.638	0.000		
19	11.518	10.200	8.940	6.407	7.391	11.676	10.710	8.008	10.630	10.403	5.251	4.632	3.787	4.968	4.386	3.521	6.359	3.128	0.000	
20	8.163	5.730	4.865	2.494	3.689	7.832	6.179	5.210	6.301	5.853	3.913	5.704	3.381	2.294	2.072	1.740	6.930	4.198	4.119	0.000