Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Genome-wide analysis of codon usage in sesame (*Sesamum indicum* L.)

Mebeaselassie Andargie [a,*], Zhu Congyi [b]

[a] *University of Goettingen, Molecular Phytopathology and Mycotoxin Research, Grisebachstrasse 6, 37077 Goettingen, Germany*

[b] *Key Laboratory of South Subtropical Fruit Biology and Genetic Resource Utilization (MOA), Guangdong Province Key Laboratory of Tropical and Subtropical Fruit Tree Research, Institute of Fruit Tree Research, Guangdong Academy of Agricultural Sciences, Guangzhou, China*

ARTICLE INFO

Keywords:
Codon usage bias
Natural selection
Optimal codon
Sesamum indicum

ABSTRACT

Sesamum indicum is an ancient oil crop grown in tropical and subtropical areas of the world. We have analyzed 23,538 coding sequences (CDS) of *S. indicum* to understand the factors shaping codon usage in this important oil crop plant. We identified eleven highly preferred codons in *S. indicum* that have AT-endings. The slope of a neutrality plot was less than one while effective number of codons (ENC) plot showed distribution above and below the standard curve. There is a significant relationship between protein length and relative synonymous codon usage (RSCU) at the primary axis while there is a weak correlation between protein length and *Nc* values. Correspondence analysis conducted on RSCU values differentiated CDS based on their GC content and their characteristic feature and showed a discrete distribution. Moreover, by determining codon usage, we found out that majority of the lignan biosynthesis related genes showed a weaker codon usage bias. These results provide insights into understanding codon evolution in sesame.

1. Introduction

Triplet codons are the basic coding units of mRNA. Evidence has shown that synonymous codons are not used randomly during translation [1, 2]. The groups of synonymous codons that encode for a specific amino acid are very well conserved over most species although a few small exceptions have been reported [3, 4]. Codon usage bias (CUB) refers to the difference in the frequency of synonymous codons in protein-coding genes from the frequencies expected if synonymous codons were selected randomly. Several factors may account for codon usage bias [5] which was observed in all forms of life. It is a complex and important phenomenon, which may have significant relevance to the understanding of genome evolution through identification of the different selective forces [6, 7, 8, 9, 10]. CUB is also critical in shaping cellular function and gene expression through its effects that ranges from RNA processing to protein translation and protein folding [11, 12, 13, 14, 15]. There are several factors that have been suggested to affect codon usage bias, while mutational bias and selection for translation efficiency are considered as the two primary evolutionary forces with varying relative contribution among species [16, 17, 18, 19, 20, 21, 22]. Mutation is responsible to generate codon diversity while reduction of codon diversity is achieved mainly through natural selection. Further factors suggested to affect CUB include compositional constraints of genes [23, 24, 25, 26], translational selection [27, 28, 29, 30], gene expression level [31, 32, 33, 34, 35, 36, 37, 38], gene length [39, 40, 41, 42], function of the gene [43], the frequency rate of recombination [27, 44], secondary structure of the protein [45, 46, 47], protein amino acid composition [48, 49, 50], the evolutionary age of the genes [51], the length of the intron [52], tRNA abundance [5, 53, 54, 55], and environmental stress [56]. Codon usage indices are used to help the tabulation and investigation of codon usage and they can reduce the codon usage data into a useful summary [7, 9, 27, 32, 36, 52, 57, 58, 59].

The advancement of genome sequencing facilitates the analysis of codon usage in many organisms, helping us to understand evolutionary forces that affect genes [5, 60]. Codon usage bias of several organisms have been analyzed so far; however, the unavailability of high-throughput sequencing of the full gene of *Sesamum indicum* hamper codon usage bias studies in *S. indicum*. *S. indicum* is an ancient oil crop grown in tropical and subtropical areas [61]. Sesame seeds contain valuable oil with anti-oxidative lignans and possess health-promoting properties [62, 63]. Here we report an analysis of codon usage in *S. indicum* and determine key factors that shape codon choice in protein-coding genes of *S. indicum*.

* Corresponding author.
E-mail address: mebhel@yahoo.com (M. Andargie).

https://doi.org/10.1016/j.heliyon.2021.e08687
Received 15 October 2021; Received in revised form 20 November 2021; Accepted 24 December 2021

2. Materials and methods

2.1. Coding sequence data

The coding sequences (CDS) datasets of *S. indicum* from the genome sequence databases of Sinbase (http://ocri-genomics.org/Sinbase) [64], and the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/) were retrieved. First, those sequences that are <300 bp were removed to minimize sampling errors. Next, coding sequences that do not possess a start and/or a stop codon as well as those having undetermined nucleotides (N) were filtered out. Later, duplicated sequences were discarded from the dataset. Finally, the 23, 538 coding sequences of *S. indicum* were used for codon usage analysis.

RSCU, which is an index to a normalized codon usage [27, 43] was calculated for each gene. It is the ratio of the actual proportional usage of a given codon to the expected proportional usage, if all the codons are used equally. Codon usage is said to be unbiased if the RSCU value is equal to 1, while a value greater than 1 indicates that a particular codon is favored. A codon with an RSCU value greater than 1.6 is considered as an overrepresented codon, and one which is having an RSCU value less than 0.6 is considered an underrepresented codon. The effective number of codons (N_C) method was used to quantify the absolute codon usage bias of a gene independent of the gene length and number of encoded amino acids [65]. The values of N_C range from 20 (when only one codon is used per amino acid) to 61 (when all codons are used in equal probability).

The GC content of first, second and third codon position (GC1, GC2 and GC3) were then calculated. GC12 is the average of GC 1 and GC 2 and was used for analysis of neutrality plots (GC12 vs GC3) [66]. The GC3s is the frequency of GC at the third codon position and it was used to better elucidate the codon usage variation and compositional constraints. Moreover, A3s, G3s, C3s, and T3s values were obtained and used to quantify the usage of each specific base at synonymous third codon positions. The codon adaptation index (CAI) was used to estimate the extent of bias toward codons that were known to be preferred in highly expressed genes. A CAI value is between 0 and 1.0, and a higher value means a likely stronger codon usage bias and a potential higher expression level [32].

2.2. Correspondence analysis (CA)

Correspondence analysis (CA) has been widely used to explore the major trends in codon usage variation among genes [67]. Correspondence analysis was performed on RSCU data to overcome the effect of biases in amino acid composition. The analysis begins with a codon usage matrix that has dimensions X (number of genes) by Y (Codon usage values). Basically, this method plots genes according to their synonymous codon usage in a 59-dimensional space (excluding Met, Trp and stop codons), then it identifies the major trends within this dataset as those axes through this multidimensional approach which account for the largest fractions of variation among these genes [67, 68, 57]. In addition, it eliminates excess noise and complex data structure by providing visual outputs [1, 69].

2.3. N_C-plot

N_C values are plotted against the GC3s values in the N_C-plot and it is used to analyze the influence of base composition on the codon usage in a genome [70]. The functional relationship that existed between N_C and GC3 values is shown by a standard curve which gives much emphasis for mutation pressure rather than selection pressure. The predicted N_C values will lie on or around the GC3 curve if the codon choice is mainly due to the G + C mutation bias. However, if the values deviate considerably below the expected GC3 curve, it signifies the presence of other factors, such as selection effects. In addition, to estimate the difference between the observed and the expected N_C values for all sesame CDS, frequency distributions of ($N_C$exp − $N_C$obs)/Ncexp were plotted.

2.4. PR2-bias plot

The parity rule 2 (PR2) is a rule of DNA composition and it is used to indicate the impact of both mutation and selection pressure on codon usage bias [29]. Plotting of AT-bias [A3/(A3+T3)] and GC-bias [G3/(G3+C3)] at the third codon position of the four-codon amino acids (i.e. Ala, Arg, Gly, Leu, Pro, Ser, Thr, and Val) is important for this analysis. When there is no deviation between mutation and selection pressure of two DNA chains, the fractional content of the four bases follows A = T and G = C (where A + T + G + C = 1) [58].

2.5. Neutrality plot

The neutrality plot (GC12-GC3) is used to estimate and characterize the relationship between GC12 and GC3 and used to determine the codon usage patterns and biases [71]. Mutation bias is assumed to be the main force shaping codon usage bias if the correlation between GC12 and GC3 is statistically significant and the slope of the regression line is close to 1. Otherwise, a slope of 0 shows the absence of directional mutation pressure [66].

2.6. Statistical analysis

To calculate the different indices of codon usage bias in *Sesamum indicum*, CodonW1.4.4 software [72] (http://codonw.sourceforge.net/), EMBOSS CUSP and CHIPS online service program were used [73]. GRAVY, Aromo, RSCU and ENC values were also calculated and correlation analyses based on Pearson's rank correlation (at a significance level of $P < 0.05$ or $P < 0.01$) were performed using the statistical software SPSS 19.0 (IBM, Chicago, IL, USA). Graphs were generated with GraphPad Prism 8.0 (GraphPad Software Inc., La Jolla, CA, USA).

3. Results

3.1. GC content variation, base composition and codon usage bias in S. indicum genome

Gene regulation as well as gene function in different organisms is mainly influenced by base composition or the proportion of guanine and cytosine bases present in the DNA molecule. As a highly variable trait, variation in GC content is observed in both synonymous and nonsynonymous sites. The synonymous codon usage pattern as well as the preference for GC and/or AT-ended codons were analyzed by the relative synonymous codon usage (RSCU) analysis of codons. In our analysis, the RSCU value of a codon greater than one indicates the codon is more frequently used whereas, RSCU value less than one indicates the codon is less frequently used in the CDS. The overall RSCU values of *S. indicum* genes further showed that 27 codons have an RSCU value greater than 1 and they were the most frequently used ones among the 59 codons. From these 27 codons, 22 codons are AT-ending codons which were used predominantly (Table 1). In addition, it has been seen that T-ending codons (14) were mostly favored as compared to A-ending codons (8) followed by five G-ending codons in the CDS of *S. indicum*. This result clearly showed that *S. indicum* genes exhibited pertinently more bias towards AT-ending codons and the genome appears to be AT rich, suggesting that the compositional constraints are the most important factor in shaping the codon usage patterns of *S. indicum*.

Based on the output of the synonymous usage of codons, five codons were identified as high-frequency codons (Table 1) of all the 59 codons after removing the three stop codons together with Methionine and tryptophan. UUG (Leu), GUU (Val), AGA, AGG (Arg) and GCU (Ala) were some of the highly preferred codons while codons such as UCG (Ser), ACG (Thr) and GCG (Ala) were particularly avoided in the *S. indicum* CDS. Moreover, the GC content value for the 23,538 genes ranged from 33.3 to 68.9 %, with an average value of 46.88 %, indicating that *S. indicum* has a high AT content.

Table 1. Codon usage of *Sesamum indicum*.

Amino acid	Codon	N	RSCU	Amino acid	Codon	N	RSCU
Phe	UUU	218,627	1.08	Ser	**UCU**	210,054	**1.43**
	UUC	187,400	0.92		UCC	134,481	0.92
Leu	UUA	108,212	0.68		UCA	185,541	1.27
	UUG	234,522	**1.48**		UCG	82,163	0.56
	CUU	215,505	**1.36**	Pro	**CCU**	174,638	**1.38**
	CUC	142,937	0.90		CCC	89,309	0.70
	CUA	90,524	0.57		**CCA**	165,595	**1.31**
	CUG	160,661	1.01		CCG	77,842	0.61
Ile	**AUU**	232,353	**1.36**	Thr	**ACU**	164,709	**1.37**
	AUC	148,610	0.87		ACC	103,609	0.86
	AUA	133,024	0.78		ACA	146,978	1.22
Met	AUG	238,739	1.00		ACG	64,844	0.54
Val	**GUU**	237,229	**1.47**	Ala	**GCU**	264,788	**1.51**
	GUC	121,122	0.75		GCC	143,209	0.82
	GUA	96,779	0.60		GCA	212,485	1.21
	GUG	190,067	1.18		GCG	79,502	0.45
Tyr	UAU	153,170	1.11	Ter	**UGA**	10,538	**1.34**
	UAC	121,840	0.89		UAA	7,089	0.90
Cys	UGU	90,992	1.01		UAG	5,910	0.75
	UGC	89,732	0.99	Trp	UGG	126,740	1.00
His	CAU	143,870	1.22	Arg	CGU	66,194	0.73
	CAC	91,674	0.78		CGC	50,368	0.56
Gln	CAA	181,110	1.00		CGA	59,332	0.66
	CAG	179,813	1.00		CGG	59,579	0.66
Asn	AAU	254,311	1.19		**AGA**	164,366	**1.82**
	AAC	174,393	0.81		**AGG**	142,731	**1.58**
Lys	AAA	270,541	0.93	Gly	GGU	179,791	1.09
	AAG	313,367	1.07		GGC	131,385	0.80
Asp	**GAU**	352,544	**1.33**		GGA	204,465	1.24
	GAC	176,211	0.67		GGG	142,337	0.87
Glu	GAA	331,043	1.03	Ser	AGU	143,087	0.98
	GAG	312,300	0.97		AGC	124,387	0.85

Note: Codons with RSCU >1.30 are shown in bold, codons with RSCU <0.80 are shown in the underlined text.

In order to further understand the nucleotide distribution, all the protein coding genes were concatenated to one sequence, which comprised 9,911,268 codons. We examined the overall GC content and GC content in the first (GC1), second (GC2), and third (GC3) codon positions and the distribution plotted (Figure 1A). A unimodal distribution was easily apparent for all GC1, GC2 and GC3 where only GC1 had higher GC content than the overall GC content. This analysis showed that there is a significant difference in the three synonymous codons. Overall, the GC content values were highest at the first codon position, followed by GC in the third and the second codon positions (GC1 > GC3 > GC2), respectively. The average value for GC1, GC2 and GC3 was 51.95 %, 41.86 % and 46.85 %, respectively (Supplementry Table 1). As expected, the peak containing high GC content genes was most pronounced in the third position. As it is clearly shown in Figure 1A, GC3 is the most important factor in *S. indicum* codon usage bias among GC, GC1, GC2, and GC3.

Besides, we examined the relationship between nucleotide content and codon usage in *S. indicum* using the effective number of codons (N_C)-plot analysis. As it is shown on Figure 1B, majority of the genes were aggregated close to the expected curve, indicating that nucleotde composition bias at the third codon position caused the observed codon bias in *S. indicum* genes. Besides, several genes with low N_C values that are located furthest away from this line were also observed, suggesting that these genes might have additional codon usage bias factors like selection forces besides the mutational bias. To get a better insight, we calculated ($N_C$exp-$N_C$obs)/$N_C$exp ratio among the observed and expected N_C values of the *S. indicum* genes (Figure 1C). The frequencies of the genes were highest when the value was within 0.0–0.05, besides the ($N_C$exp-N$_C$obs)/$N_C$exp value ranges in between −0.05–0.25 for most of the genes, which indicates that the observed N_C values are not identical to the expected N_C values according to their GC3s. This result provided more evidence for the existence of other factors that affected the *S. indicum* codon usage bias.

Furthermore, the association between purines (A and G) and pyrimidines (C and T) at the third codon position was carried out using Parity Rule 2 (*PR2*) bias plot in order to further elucidate the impact of selection and mutation pressure on the CUB of *S. indicum* genes. If the codon usage bias is occurred mainly due to mutation bias, G and C (A and T) should be used proportionally at the third codon position. On the other hand, if natural selection dominates, it would not necessarily cause proportional usage of G and C (A and T). In this analysis, we have noticed that majority of the genes were distributed at the third and fourth quadrat of the *PR2*-plot (Figure 1D), implying that there exists a codon usage imbalance between A + T and G + C at the third codon position of the *S. indicum* genes.

To further investigate the driving forces that are operating behind and to explore the magnitude of mutational pressure against the effect of natural selection in *S. indicum* genes, neutrality plot analysis was conducted (Figure 1E). The neutrality plot (GC12 vs GC3s) revealed that the *S. indicum* genome had a wide range of GC3 distributions and there exists a non-linear relationship between GC3 and codon usage bias in *S. indicum* CDS. In addition, the slope of the estimated equation revealed that

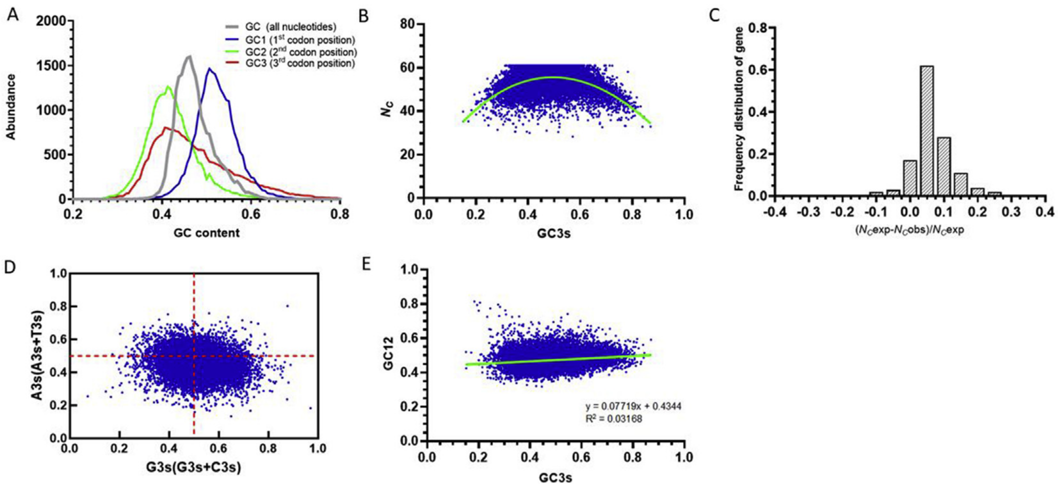

Figure 1. Analysis of codon usage in *Sesamum indicum*. (A) The distribution of GC contents at the three codon positions in *S. indicum* genes, (B) Distribution of effective number of codons (N_C) and GC3s of *S. indicum* genes. Individual genes are indicated by dots while the standard curve represents the expected N_C under random codon usage, (C) Frequency distribution of effective number of codons (N_C) ratio, (D) Parity Rule 2 (*PR2*) bias plot analysis. Genes are plotted based on their GC bias [G3/(G3+C3)] and AU bias [A3/(A3+T3)] in the third codon position, (E) Neutrality plot analysis (GC12 vs GC3s) of *S. indicum*. GC12 is the average value of Guanine and Cytosine content in the first (GC1) and second (GC2) position of the codons, whereas GC3 is the Guanine and Cytosine content at the third codon position.

mutation pressure accounted only for 7 % while selection constraints accounted for 93 % of the effects observed indicating that GC12 content was affected by mutation pressure and natural selection at a ratio of 0.7/0.93 = 0.75. In general, the above-mentioned findings suggested that natural selection might have played a major role while mutation pressure might have played a minor role in the evolution of codon usage bias of the *S. indicum* genes.

3.2. Impact of nucleotide composition in affecting codon bias

To investigate synonymous codon usage variation among *S. indicum* genes, correspondence analysis (COA) was performed on the RSCU values of each codon. Based on our COA analysis of the RSCU values, the first principal axis described 15.57 % of the variation while the second axis accounted for only 5.26 % of the total variation (Figure 2A). This result confirmed that the first axis reflects a significant trend that explain differences in codon usage among the *S. indicum* genes. As it is shown on Figure 2B, *S. indicum* genes showed a broad distribution along the horizontal axis, and detailed analysis indicated that *S. indicum* genes are categorized in to three distinct groups. Genes that have a GC content which is less than 45 % were partitioned at the right side of axis 1 while genes having GC ≥ 60 % were located mainly at the left side of axis 1. Moreover, genes having a GC content of 45 %–60 % were distributed mainly at the center of the plot. Interestingly, those genes located at the left side of axis 1 tend to be the genes with stronger codon usage bias as it is assessed using their N_C values while the ones that are located at the other extreme of axis 1 are expected to be expressed at low levels. Thus, these results clearly showed that correspondence analysis differentiates genes according to their codon usage differences in addition to showing and it also showed the effects of nucleotide composition at each codon.

In addition, aromatic genes and hydrophobic genes that has a value of ≥0.15, >0.3, respectively as well as ribosomal and other genes were selected to distinguish the pattern of codon usage among the *S. indicum* genes. The distribution of these genes was marked along axis 1 and axis 2 based on the correspondence analysis (Figure 2C). Most of the ribosomal genes were gathered at the right side of axis 1, while other genes were located mainly on the left side of axis 1. Besides, both aromatic and hydrophobic genes were concentrated mainly at the upper half region of axis 1. Altogether, these results showed that the codon usage bias of these genes might be influenced by base composition under mutation. Besides, the effect of natural selection on codon usage bias of these genes to some extent is clearly observed since the genes showed a discrete distribution.

Nucleotide composition is one of the major factors for the mutational pressure in codon usage bias of genes. So, correlation analysis was performed to ascertain the association between the two principal axes (axis 1 and axis 2) and nucleotide composition (Table 2). A significant positive correlation ($P < 0.0001$) was observed between axis 1 and A, T, C, G, A3, C3, and G3 while a significant negative correlation ($P < 0.0001$) was

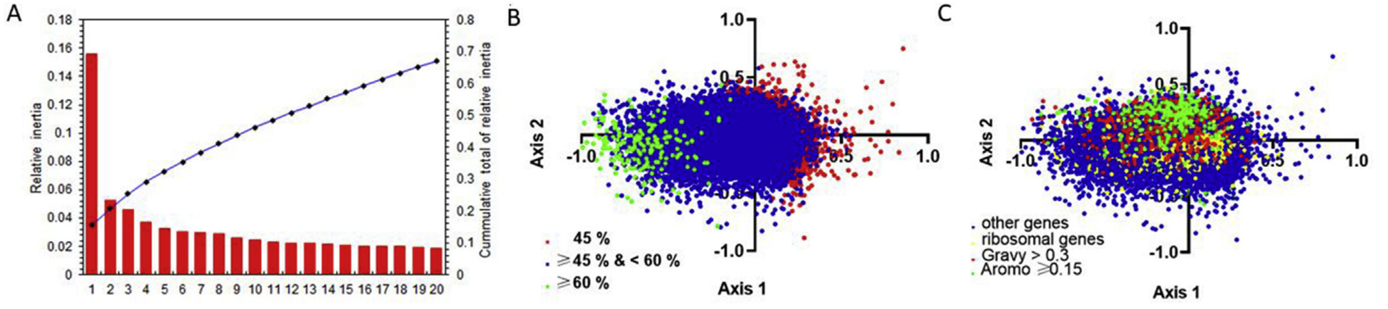

Figure 2. Effects of nucleotide composition on codon usage bias. (A) The relative first 20 factors from correspondence analysis according to their amino acid proportions. The line represents the cumulative total of the inertia explained by the first 20 axis. (B) Correspondence analysis of RSCU values of genes. Red, blue and green dots indicate genes with guanine and cytosine contents that are <45 %, ≥45 %, & < 60 % and ≥60 %, respectively. (C) Correspondence analysis on RSCU values of codons. Blue dots, yellow dots, red dots and green dots indicate other genes, ribosomal genes, genes with a Gravy value >0.3 and genes with Aromo value ≥0.15, respectively.

Table 2. Correlation analysis of axis 1 and axis 2 with overall nucleotide composition and effective number of codons.

	A	T	C	G	A3	T3	C3	G3	GC%	GC1%	GC2%	GC3%	N_c
Axis 1	0.3630	0.3373	0.1798	0.2733	0.4106	**−0.00758**	0.4038	0.2047	−0.7859	−0.1717	−0.3019	−0.8551	−0.08383
P	0.000	0.000	0.000	0.000	0.000	**0.2451**	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Axis 2	0.04344	0.1770	0.1351	0.04594	0.08181	0.1854	0.1329	−0.01512	−0.2078	−0.2643	0.07242	−0.2071	0.05478
P	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.0204	0.000	0.000	0.000	0.000	0.000

A, T, C, G = frequency of each individual base, A3, T3, C3 and G3 = frequency of each individual base A, T, C and G at the third position of codons, GC = total guanine-cytosine content of the entire gene, GC1 = GC content in the first position of codons, GC2 = GC content in the second position of codons, GC3 = GC content in the third position of codons, and N_C = effective codon number. N_C = effective codon number. Significant difference at $P < 0.0001$. Bold value indicates that there is no significant correlation at $P < 0.0001$.

observed between axis 1 and GC %, GC1 %, GC2 %, and GC3 %. However, axis 1 had no significant correlation with T3. Similarly, axis 2 showed a significant positive correlation with A, T, C, G, A3, T3, C3, and GC2 % while it showed a significant negative correlation with G3, GC %, GC1 % and GC3 %. Besides, a negative and positive correlation was obtained between *Nc* and axis 1 as well as axis 2, respectively suggesting that mutational pressure which arises from base composition at least partly might influence the codon usage bias of *S. indicum* genes.

Moreover, correlation analysis was performed between axis 1 as well as axis 2 with indices like protein length, Aromo, and GRAVY values, which are important factors to govern natural selection (Table 3). As it is clearly shown on Table 3 below, there was a significant positive correlation between axis 1 and protein length while there was a significant negative correlation between axis 1 and Aromo as well as GRAVY respectively. Similarly, axis 2 had a significant positive correlation with both Aromo and GRAVY values while it showed no significant correlation with protein length. These results suggested that codon usage bias in *S. indicum* genes might also have emerged from the effect of natural selection.

3.3. Effects of gene length on codon usage bias

We have found a significant positive correlation between protein length and RSCU primary axis (axis 1) (r = 0.324, $P < 0.001$); however, there was a weak linear correlation between protein length and *Nc* (r = 0.0093, $P = 0.1543$). It is worth to note the presence of two different patterns that were observed when we did our analysis on the variation that existed among proteins: i.e short protein products had the greatest variation in *Nc* values and protein length showed a significant variation in relation to higher *Nc* values (Figure 3). Most of the data points were distributed between *Nc* values of 48 and 58. This result showed that length of proteins contributed for *S. indicum* codon usage bias which arises due to selection constraints.

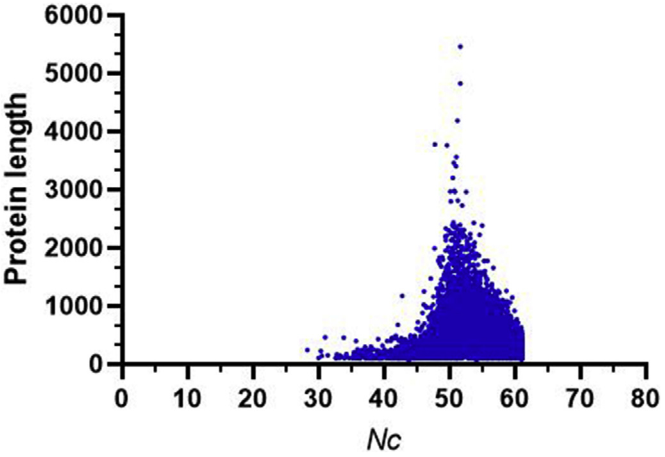

Figure 3. Plot of protein length versus the N_C value variation.

3.4. Effect of hydrophobicity and aromaticity of encoded proteins on codon usage bias

Correlation analysis among, Aromo, GRAVY and N_C were determined in order to investigate the potential effect of physical and chemical properties of the encoded proteins on *S. indicum* codon usage bias. As it is shown in Table 4 below, GRAVY and Aromo values of the amino acids were positively correlated with *Nc* for each protein. This result confirms that both hydrophobicity and aromaticity influence the codon usage bias in *S. indicum.*

Table 3. Correlation analysis of axis 1 and axis 2 with Protein length, Aromo and GRAVY values.

	Protein length	Aromo	GRAVY
Axis 1	0.3239	−0.03343	−0.06806
P	0.000	0.000	0.000
Axis 2	**0.01076**	0.05092	0.1123
P	**0.0989**	0.000	0.000

Aromo = frequency of aromatic amino acids, GRAVY = General average hydrophobicity Significant difference at $P < 0.0001$. Bold value indicates that there is no significant correlation at $P < 0.0001$.

3.5. Lignan biosynthesis related genes of S. indicum and their codon usage bias

We have taken all the seven known lignan biosynthesis related gene CDS and analyzed their codon usage patterns. To achieve our aim, GC12/GC3 ratios and *Nc* values were used to calculate the codon usage pattern of the lignan biosynthesis related genes. Later, the obtained values were compared to the average values of all *S. indicum* coding sequences. The results showed that from the seven lignan biosynthesis related genes CYP81Q1, CYP81Q3 and UGT71A9 have plotted data points which are above the average *Nc* values while CYP92B14 and UGT94AG1 have plotted data points which showed a higher value compared to the average GC12/GC3 ratios, indicating a weaker codon usage bias when it is compared to the average (Figure 4A). Furthermore, we found out that CYP92B14, UGT94D1 and UGT94AG1 had lower *Nc* values while CYP81Q3, UGT71A9, UGT94D1 and UGT94AA2 had a much lower GC12/GC3 ratio below the average. Moreover, correspondence analysis showed that the position of the AT ended codons are mainly located at the right side of axis1 while the GC ending codons concentrate mainly at

Table 4. Correlation analysis of *Nc* with GRAVY and Aromo values.

	GRAVY	Aromo
Nc	0.07161	0.1132
P	0.000	0.000

GRAVY = General average hydrophobicity, Aromo = frequency of aromatic amino acids, *Nc* = effective codon number Significant difference at $P < 0.0001$.

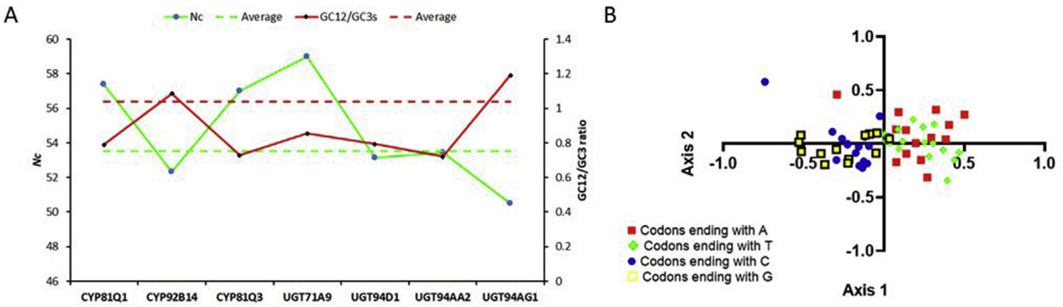

Figure 4. Codon usage analysis in *S. indicum* lignan biosynthesis related genes. (A) Comparisons of GC12/GC3 ratio and N_C values among different lignan biosynthesis related genes. The average N_C value is 53.54 while the general GC12/GC3 value is 1.04. (B) Correspondence analysis of the synonymous codon usage of *S. indicum* lignan biosynthesis related genes. Different base ended codons were marked in the figure, where the red, green, blue, and yellow colors refer to codons ending with A, T, C, G respectively.

the left side of axis 1 (Figure 4B), indicating that the base composition for mutation bias might correlate to the codon bias.

4. Discussion

The overall codon usage pattern among the 23,538 of *Sesamum indicum* CDS was analyzed in this study. Usually RSCU values are used as a measurement of codon usage bias and the RSCU bias differs accordingly among various species and genes [74]. The possible causes of RSCU bias have been investigated in the genomes of numerous living organisms, for example, in *Zea mays, Arabidopsis thaliana*, cotton, and so others [10, 75, 59]. Accordingly, in the current study, we have found that the A/T-ending codons were the more frequent and preferred ones compared to the G/C-ending codons in *S. indicum* genome. From the 59 synonymous codons, 27 codons had RSCU value which is greater than 1 and out of these 22 were AT-ending while only five were GC-ending codons. This showed that the genes had little or no bias at all towards the GC-ending codons in *S. indicum*.

In order to ascertain the potential effect of compositional constraints on codon usage, first the nucleotide compositions of the above-mentioned *S. indicum* CDS were determined. It was found that the contents of GC1, GC2, GC3 and the average content of GC at three positions were less than 50 %, indicating that the *S. indicum* genome tended to use A/T bases and AT-ending codons more frequently. Moreover, the *S. indicum* coding sequences were found to be AT rich and compared to A, T is the most frequently preferred nucleotide. A random synonymous codon choice whereby the Guanine-Cytosine on the one hand and the Adenine-Thymine on the other hand would be used proportionally among the degenerate codon groups in a gene or genome, if mutation occurs at the third codon position neutrally [35]. The current study showed the unbalanced usage of AT and GC at the third codon position of *S. indicum* genes. This in turn supports the influence of mutation pressure. Similar observations were reported previously on comparative analysis of codon usage in the *Poaceae* family [76], the *Asteraceae* family [9], the *Solanum* species [77], *Euphorbiaceae* species [78], and *Paeonia lactiflora* [79]. Here, it is worth to note that the GC value of *S. indicum* (46.88 %) was found to be higher with those reported for dicots (42–44 %) [58]. In this study, the variation among the codon usage patterns of different *S. indicum* genes is also reflected in the effective *Nc* values, which range from 61 to the very low value of 28.3, corresponding to an extremely biased codon usage. [65] suggested plotting *Nc* against GC3s as a means of characterizing codon usage variation among genes. In line with this, if codon usage is affected only by GC3 content, the *Nc* values lie just over the expected *Nc* curve, which indicates mutational pressure [80]. Even though, the points for most genes follow this trend in this study, but there are some genes that lie below the expected curve corresponding to the *Nc* values, which indicates the dominant role played by selection pressure. Furthermore, the PR2 plot as well as the neutrality plot analysis showed that the *S. indicum* genes did not use GC and AT equally and the slope (0.077) of the regression line which is constructed by GC12 vs GC3 was found to be close to zero, respectively. This unequal usage which is observed in the degenerate codon positions as well as the nearly horizontal line of the neutrality plot analysis gives evidence that natural selection has played a major role in shaping the codon usage bias in *Sesamum indicum*. Similarly, in *Oryza sativa*, *Zea mays*, *Arabidopsis thaliana*, *Triticum aestivum* [81], *Silene latifolia* [82] and *Epichloë festucae* [83] natural selection plays a much important role in forming the codon usage bias. Generally, in order to determine codon usage bias in different organisms, it is important to maintain a balance between mutation pressure and natural selection [77, 84, 85].

According to the obtained correspondence analysis result, codon usage variation might be one of the factors that contributed as a driving force for CUB of *S. indicum* genes. To this end, we further performed correspondence analysis of RSCU values to ascertain the major trends in the variation of codon usage. The observed high degree of correlation between positions of genes along axis 1 and GC3s as well as C3s in comparison to G3s can help to conclude that there is one major source of synonymous codon usage variation among these *S. indicum* genes, visible in the GC3s content. However, when the first axis is observed, it was able to explain only a partial amount of variation (15.57 %) of codon usage among these genes while axis 2 accounted for 5.26 %. So, from this investigation, it is possible to conclude that besides mutational bias and natural selection, there might be several other factors that are responsible at least partially to determine codon usage bias and variation in *S. indicum* genes.

Correspondence analysis also revealed a strong positive correlation between protein length and axis 1 in the current study; however, protein length and *Nc* showed a weak linear correlation. Since the cost of translating a protein is directly proportional to the length of a protein, there might be a greater pressure for the selection of the most accurate codons in longer genes compared to shorter genes in order to avoid missense errors. In eukaryotic organisms, it has been argued that selection may act to decrease the length of highly expressed genes [28]. Compared to the shorter proteins, a weak codon usage bias is displayed by longer proteins; however, shorter proteins did not particularly affect the codon usage. This indicates that the codon usage pattern in *S. indicum* might be shaped by other selective forces.

Previous studies have showed that there are other natural selection driven factors, which cause codon usage bias, such as Grand average of hydropathicity (GRAVY) and aromaticity of amino acid usage (AROMO) values [86]. The significant positive correlations of GRAVY and AROMO values supports the finding that both values of proteins, which are encoded by the coding sequences could be associated with the codon usage bias of *S. indicum* genes. Previous studies have found significant positive and negative correlations between GRAVY and AROMO values and codon usage bias in a variety of organisms, such as *Pisum* species [22], *Ginkgo biloba* [77], *Oncidium Gower Ramsey* [87], *Epichloë festucae* [9]. Finally, in the current study, I investigated the codon usage bias in the seven known lignan biosynthesis related genes and observed substantial differences in the codon usage among the lignan biosynthesis related genes.

5. Conclusion

In summary, the current study revealed the pattern of codon usage bias in *Sesamum indicum* genome in association with the different factors that are responsible to bring the bias. Based on the current investigation, codon usage bias in *S. indicum* appears to be a combined effort of several factors like nucleotide composition, mutation bias, natural selection, protein length (at least partially), aromaticity and hydropathicity.

Declarations

Author contribution statement

Mebeaselassie Andargie; Zhu Congyi: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

Funding statement

Mebeaselassie Andargie was supported by the Alexander von Humboldt Foundation through Georg Förster Research Fellowship.

Data availability statement

Data included in article/supplementary material/referenced in article.

Declaration of interests statement

The authors declare no conflict of interest.

Additional information

Supplementary content related to this article has been published online at https://doi.org/10.1016/j.heliyon.2021.e08687.

Appendix

Supplementary Table 1. Means and standard deviations of several index numbers from 23,538 genes in *Sesamum indicum*.

Class	Genes	Codons	GC all (%)	GC1 (%)	GC2 (%)	GC3 (%)	GC3s (%)	T3s (%)	C3s (%)	A3s (%)	G3s (%)	Gravy	Aromo	ENC	CAI
Total	23,538	9,911,268	46.89 ± 4.39	51.95 ± 6.22	41.86 ± 4.68	46.85 ± 4.19	44.86 ± 9.37	37.47 ± 7.59	27.11 ± 7.87	30.49 ± 6.68	30.21 ± 6.81	−0.31 ± 0.37	0.08 ± 0.03	53.54 ± 4.53	0.20 ± 0.03

GC all = total guanine-cytosine content of the entire gene, GC1 = GC content at the first, GC2 = GC content at the second, GC3 = GC content at the third codon positions, GC3s = proportion of GC nucleotides at the third (variable) coding position of synonymous codons, T3s, C3s, A3s and G3s = frequency of each individual base A, T, G and C at the third position of codons, Gravy = General average hydropathicity, Aromo = frequency of aromatic amino acids, ENC = effective codon number, CAI = Codon adaptation index.

References

[1] R.C. Grantham, M. Gautier, R. Gouy, A. Mercier, A. Pave, Codon catalog usage and the genome hypothesis, Nucleic Acids Res. 8 (1980) r49–62.

[2] T. Ikemura, Codon usage and tRNA content in unicellular and multicellular organisms, Mol. Biol. Evol. 2 (1985) 13–34.

[3] S. Osawa, T.H. Jukes, K. Watanabe, A. Muto, Recent evidence for evolution of the genetic code, Microbiol. Rev. 56 (1992) 229–264.

[4] M.A. Santos, G. Moura, S.E. Massey, M.F. Tuite, Driving change: the evolution of alternative genetic codes, Trends Genet. 20 (2004) 95–102.

[5] J.B. Plotkin, G. Kudla, Synonymous but not the same: the causes and consequences of codon bias, Nat. Rev. Genet. 12 (2011) 32–42.

[6] P.M. Sharp, L.B. Emery, K. Zeng, Forces that influence the evolution of codon bias, Philos. Trans. R. Soc. 365 (2010) 1203–1212.

[7] H. Chiapello, F. Fisacek, M. Caboche, A. Henaut, Codon usage and gene function are related in sequences of *Arabidopsis thaliana*, Gene 209 (1998) GC1–GC38.

[8] M. Zhou, X. Li, Analysis of synonymous codon usage patterns in different plant mitochondrial genomes, Mol. Biol. Rep. 36 (8) (2009) 2039–2046.

[9] S. Qiu, R. Bergero, K. Zeng, D. Charlesworth, Patterns of codon usage bias in *Silene latifolia*, Mol. Biol. Evol. 28 (1) (2011) 771–780.

[10] L.Y. Wang, H.X. Xing, Y.C. Yuan, X.L. Wang, M. Saeed, J.C. Tao, W. Feng, G.H. Zhang, X.L. Song, X.Z. Sun, Genome-wide analysis of codon usage bias in four sequenced cotton species, PLoS One 13 (2018), e0194372.

[11] M. dos Reis, L. Wernisch, R. Savva, Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome, Nucleic Acids Res. 31 (2003) 6976–6985.

[12] Q. Liu, S. Dou, Z. Ji, Q. Xue, Synonymous codon usage and gene function are strongly related in Oryza sativa, Biosystems 80 (2) (2005) 123–131.

[13] T. Tuller, A. Carmi, K. Vestsigian, S. Navon, Y. Dorfan, J. Zaborske, T. Pan, O. Dahan, I. Furman, Y. Pilpel, An evolutionarily conserved mechanism for controlling the efficiency of protein translation, Cell 141 (2010) 344–354.

[14] C. Pop, S. Rouskin, N.T. Ingolia, L. Han, E.M. Phizicky, J.S. Weissman, D. Koller, Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation, Mol. Syst. Biol. 10 (2014) 770.

[15] T.E.F. Quax, N.J. Claassens, D. Söll, J. van der Oost, Codon bias as a means to fine-tune gene expression, Mol. Cell. 59 (2015) 149–161.

[16] R.M. Kliman, J. Hey, The effects of mutation and natural selection on codon bias in the genes of *Drosophila*, Genetics 137 (1994) 1049–1056.

[17] S. Kanaya, Y. Yamada, Y. Kudo, T. Ikemura, Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis, Gene 238 (1999) 143–155.

[18] M.C. Angellotti, S.B. Bhuiyan, G. Chen, X.F. Wan, CodonO: codon usage bias analysis within and across genomes, Nucleic Acids Res. 35 (2007) W132–W136.

[19] S. Subramanian, Nearly neutrality and the evolution of codon usage bias in eukaryotic genomes, Genetics 178 (2008) 2429–2432.

[20] Q. Liu, H. Hu, H. Wang, Mutational bias is the driving force for shaping the synonymous codon usage pattern of alternatively spliced genes in rice (*Oryza sativa* L.), Mol. Genet. Genom. 290 (2) (2015) 649–660.

[21] R. Prabha, D.P. Singh, S. Sinha, K. Ahmad, A. Rai, Genome-wide comparative analysis of codon usage bias and codon context patterns among cyanobacterial genomes, Marine Genomics 32 (2017) 31–39.

[22] D. Bhattacharyya, A. Uddin, S. Das, S. Chakraborty, Mutation pressure and natural selection on codon usage in chloroplast genes of two species in *Pisum l.* (Fabaceae: Faboideae), Mitochondrial. DNA A DNA Mapp. Seq. Anal. 30 (4) (2019) 664–673.

[23] G. Bernardi, Compositional constraints and genome evolution, J. Mol. Evol. 24 (1986) 1–11.

[24] S. Osawa, T. Ohama, F. Yamao, A. Muto, T.H. Jukes, H. Ozeki, K. Umesono, Directional mutation pressure and transfer RNA in choice of the third nucleotide of synonymous two-codon sets, Proc. Natl. Acad. Sci. U.S.A. 85 (4) (1988) 1124–1128.

[25] N. Sueoka, Y. Kawanishi, DNA G+C content of the third codon position and codon usage biases of human genes, Gene 261 (2000) 53–62.

[26] G. Marais, D. Mouchiroud, L. Duret, Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes, Proc. Natl. Acad. Sci. U.S.A. 98 (2001) 5688–5692.

[27] P.M. Sharp, W.H. Li, An evolutionary perspective on synonymous codon usage in unicellular organisms, J. Mol. Evol. 24 (1986) 28–38.

[28] E.N. Moriyama, J.R. Powell, Codon usage bias and tRNA abundance in Drosophila, J. Mol. Evol. 45 (1997) 514–523.

[29] N. Sueoka, 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 (1) (1999) 53–58.

[30] P.K. Ingvarsson, Gene expression and protein length influence codon usage and rates of sequence evolution in *Populus tremula*, Mol. Biol. Evol. 24 (2007) 836–844.

[31] W. Bains, Codon distribution in vertebrate genes may be used to predict gene length, J. Mol. Biol. 197 (1987) 379–388.

[32] P.M. Sharp, W.H. Li, The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias, Mol. Biol. Evol. 4 (1987) 222–230.

[33] L. Duret, D. Mouchiroud, Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis, Proc. Natl. Acad. Sci. U.S.A. 96 (1999) 4482–4487.

[34] J. Hey, R.M. Kliman, Interactions between natural selection, recombination and gene density in the genes of Drosophila, Genetics 160 (2002) 595–608.

[35] W.J. Zhang, J. Zhou, Z.F. Li, L. Wang, X. Gu, Y. Zhong, Comparative analysis of codon usage patterns among mitochondrion, chloroplast and nuclear genes in *Triticum aestivum* L, J. Integr. Plant Biol. 49 (2) (2007) 246–254.

[36] S. Chakraborty, D. Nag, T.H. Mazumder, A. Uddin, Codon usage pattern and prediction of gene expression level in Bungarus species, Gene 604 (2017) 48–60.

[37] M.P. Victor, D. Acharya, T. Begum, T.C. Ghosh, The optimization of mRNA expression level by its intrinsic properties-insights from codon usage pattern and structural stability of mRNA, Genomics 111 (6) (2019) 1292–1297.

[38] Z. Chen, J. Zhao, J. Qiao, W. Li, J. Li, R. Xu, H. Wang, Z. Liu, B. Xing, J.F. Wendel, C.E. Grover, Comparative analysis of codon usage between *Gossypium hirsutum* and *G. barbadense* mitochondrial genomes, Mitochondrial DNA Part B 56 (3) (2020) 2500–2506.

[39] A. Eyre-Walker, Synonymous codon bias is related to gene length in *Escherichia coli*: selection for translational accuracy, Mol. Biol. Evol. 13 (1996) 864–872.

[40] E. Moriyama, J. Powell, Gene length and codon usage bias in *Drosophila melanogaster, Saccharomyces cerevisiae* and *Escherichia coli*, Nucleic Acids Res. 26 (1998) 3188–3193.

[41] J.M. Comeron, M. Kreitman, M. Aguade, Natural selection on synonymous sites is correlated with gene length and recombination in Drosophila, Genetics 151 (1999) 239–249.

[42] R.M. Kliman, J. Hey, Hill-robertson interference in *Drosophila melanogaster*: reply to Marais, Mouchiroud and Duret, Genet. Res. 81 (2003) 89–90.

[43] P.M. Sharp, T.M. Tuohy, K.R. Mosurski, Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes, Nucleic Acids Res. 14 (13) (1986) 5125–5143.

[44] T. Xie, D. Ding, X. Tao, D. Dafu, The relationship between synonymous codon usage and protein structure, FEBS Lett. 434 (1998) 93–96.

[45] J.R. Lobry, C. Gautier, Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes, Nucleic Acids Res. 22 (1994) 3174–3180.

[46] G. D'Onofrio, T.C. Ghosh, G. Bernardi, The base composition of the human genes is correlated with the secondary structures of the encoded proteins, Gene 300 (2002) 179–187.

[47] M. Oresic, M. Dehn, D. Korenblum, D. Shalloway, Tracing specific synonymous codon-secondary structure correlations through evolution, J. Mol. Evol. 56 (2003) 473–484.

[48] H. Romero, A. Zavala, H. Musto, Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern of selective forces, Nucleic Acids Res. 28 (2000) 2084–2090.

[49] C. Rispe, F. Delmotte, R.C. van Ham, A. Moya, Mutational and selective pressures on codon and amino acid usage in Buchnera, endosymbiotic bacteria of aphids, Genome Res. 14 (2004) 44–52.

[50] Y. Prat, M. Fromer, N. Linial, M. Linial, Codon usage is associated with the evolutionary age of genes in metazoan genomes, BMC Evol. Biol. 9 (2009) 285.

[51] A.E. Vinogradov, Intron length and codon usage, J. Mol. Evol. 52 (2001) 2–5.

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

[53] T. Ikemura, Correlation between the abundance of yeast transfer RNAs and the occurrence of respective codons in protein genes, J. Mol. Biol. 158 (1982) 573–597.

[54] M. Bulmer, Coevolution of codon usage and transfer RNA abundance, Nature 325 (1987) 728–730.

[55] H. Goodarzi, N. Torabi, H.S. Najafabadi, M. Archetti, Amino acid and codon usage profiles: adaptive changes in the frequency of amino acids and codons, Gene 407 (2008) 30–41.

[56] S. Vicario, E.N. Moriyama, J.R. Powell, Codon usage in twelve species of *Drosophila*, BMC Evol. Biol. 7 (2007) 226.

[57] Z. Wang, B. Xu, B. Li, Q. Zhou, G. Wang, X. Jiang, C. Wang, Z. Xu, Comparative analysis of codon usage patterns in chloroplast genomes of six *Euphorbiaceae* species, PeerJ 6 (8) (2020), e8251.

[58] F. Wright, The effective number of codons used in a gene, Gene 87 (1990) 23–29.

[59] H.M. Liu, R. He, H.Y. Zhang, Y.B. Huang, M.L. Tian, J.J. Zhang, Analysis of SCU in *Zea mays*, Mol. Biol. Rep. 37 (2010) 677–684.

[60] Wendel, J. F.; Greilhuber, J.; Doležel, J.; Leitch, I. J.; Flagel, L. E.; Blackman, B. K. The first ten years of plant genome sequencing and prospects for the next decade. Plant Genome Diversity, eds Wendel JF, Greilhuber J, Doležel J, Leitch IJ (Springer, Vienna), Vol 1, pp 1–15

[61] Bedigian, D. Sesame: the genus sesamum. Ed Bedigian D. CRC Press, Taylor & Francis Group.

[62] Y. Fukuda, M. Nagata, T. Osawa, M. Namik, Chemical aspects of the antioxidative activity of roasted sesame seed oil, and the effect of using the oil for frying, Agric. Biol. Chem. 50 (1986) 857–862.

[63] A.A. Moazzami, S.L. Haese, A. Kamal-Eldin, Lignan contents in sesame seeds and products, Eur. J. Lipid Sci. Technol. 109 (2007) 1022–1027.

[64] L. Wang, J. Yu, D. Li, X. Zhang, Sinbase: an integrated database to study genomics, genetics and comparative genomics in *Sesamum indicum*, Plant Cell Physiol. 56 (2015) e2.

[65] R. Hershberg, D.A. Petrov, General rules for optimal codon choice, PLoS Genet. (2009), e1000556.

[66] Y. Zhang, X. Nie, X. Jia, C. Zhao, S.S. Biradar, L. Wang, X. Du, S. Weining, Analysis of codon usage patterns of the chloroplast genomes in the *Poaceae* family, Aust. J. Bot. 60 (2012) 461–470.

[67] X. Nie, P. Deng, K. Feng, P. Liu, X. Du, F.M. You, S. Weining, Comparative analysis of codon usage patterns in chloroplast genomes of the *Asteraceae* family, Plant Mol. Biol. Rep. 32 (2014) 828–840.

[68] R. Zhang, L. Zhang, W. Wang, Z. Zhang, H. Du, Z. Qu, X.Q. Li, H. Xiang, Differences in codon usage bias between photosynthesis-related genes and genetic system-related genes of chloroplast genomes in cultivated and wild solanum species, Int. J. Mol. Sci. 19 (2018) e3142.

[69] Y. Wu, D. Zhao, J. Tao, Analysis of codon usage patterns in herbaceous peony (*Paeonia lactiflora* Pall.) based on transcriptome data, Genes 6 (2015) 1125–1139.

[70] R. Singh, R. Ming, Q. Yu, Comparative analysis of GC content variations in plant genomes, Trop. Plant Biol. 9 (2016) 136–149.

[71] B. He, H. Dong, C. Jiang, F. Cao, S. Tao, L.A. Xu, Analysis of codon usage patterns in *Ginkgo biloba* reveals codon usage tendency from A/U-ending to G/C-ending, Sci. Rep. 6 (2016) 35927.

[72] J.F. Peden, Analysis of codon usage, University of Nottingham, 2000.

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

[74] Q. Liu, Q. Xue, Comparative studies on codon usage pattern of chloroplasts and their host nuclear genes in four plant species, J. Genet. 84 (2005) 55–62.

[75] S. Karumathil, N.T. Raveendran, D. Ganesh, N.S. Kumar, R.R. Nair, V.R. Dirisala, Evolution of SCU bias in west African and central African strains of monkeypox virus, Evol. Bioinf. Online 14 (2018) 1–22.

[76] S. Qiu, K. Zeng, T. Slotte, S. Wright, D. Charlesworth, Reduced efficacy of natural selection on codon usage bias in selfing Arabidopsis and Capsella species, Genome Biol. Evol. 3 (2011) 868–880.

[77] X. Li, H. Song, Y. Kuang, S. Chen, P. Tian, C. Li, Z. Nan, Genome-wide analysis of codon usage bias in *Epichloë festucae*, Int. J. Mol. Sci. 17 (2016) 1138.

[78] L.A. Shackelton, C.R. Parrish, E.C. Holmes, Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses, J. Mol. Evol. 62 (2006) 51–563.

[79] S. Yengkhom, Arif. Uddin, S. Chakraborty, Deciphering codon usage patterns and evolutionary forces in chloroplast genes of *Camellia sinensis* var. *assamica* and *Camellia sinensis* var. *sinensis* in comparison to *Camellia pubicosta*, J. Integrat. Agric. 18 (2019) 2771–2785.

[80] N. Sueoka, Directional mutation pressure and neutral molecular evolution, Proc. Natl. Acad. Sci. U.S.A. 85 (1988) 2653–2657.

[81] P.M. Sharp, W.H. Li, The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications, Nucleic Acids Res. 15 (1987) 1281–1295.

[82] M.J. Greenacre, Theory and Applications of Correspondence Analysis, Academic Press, London, 1984.

[83] Q. Liu, Y. Feng, X. Zhao, H. Dong, Q. Xue, Synonymous codon usage bias in *Oriza sativa*, Plant Sci. 167 (2004) 101–105.

[84] H.C. Wang, D.A. Hickey, Rapid divergence of codon usage patterns within the rice genome, BMC Evol. Biol. 7 (Suppl. 1) (2007) S6.

[85] D.L. Hartl, E.N. Moriyama, S.A. Sawyer, Selection intensity for codon bias, Genetics 138 (1994) 227–234.

[86] P.M. Sharp, K.M. Devine, Codon usage and gene expression level in *Dictyostelium discoideum*: highly expressed genes do prefer optimal codons, Nucleic Acids Res. 17 (1989) 5029–5039.

[87] N. Sueoka, Intrastrand parity rules of DNA base composition and usage biases of synonymous codons, J. Mol. Evol. 40 (3) (1995) 318–325.