Journal of Genetics, Vol. 97, No. 1, March 2018, pp. 97–107
https://doi.org/10.1007/s12041-018-0892-1

RESEARCH ARTICLE

Codon usage *vis-a-vis* start and stop codon context analysis of three dicot species

PROSENJIT PAUL, ARUP KUMAR MALAKAR and SUPRIYO CHAKRABORTY*

Department of Biotechnology, Assam University, Silchar 788 011, India
*For correspondence. E-mail: supriyoch2008@gmail.com.

Received 17 February 2017; revised 13 May 2017; accepted 16 June 2017; published online 20 February 2018

Abstract. To understand the variation in genomic composition and its effect on codon usage, we performed the comparative analysis of codon usage and nucleotide usage in the genes of three dicots, *Glycine max*, *Arabidopsis thaliana* and *Medicago truncatula*. The dicot genes were found to be A/T rich and have predominantly A-ending and/or T-ending codons. GC3s directly mimic the usage pattern of global GC content. Relative synonymous codon usage analysis suggests that the high usage frequency of A/T over G/C mononucleotide containing codons in AT-rich dicot genome is due to compositional constraint as a factor of codon usage bias. Odds ratio analysis identified the dinucleotides TpG, TpC, GpA, CpA and CpT as over-represented, where, CpG and TpA as under-represented dinucleotides. The results of (NcExp−NcObs)/NcExp plot suggests that selection pressure other than mutation played a significant role in influencing the pattern of codon usage in these dicots. PR2 analysis revealed the significant role of selection pressure on codon usage. Analysis of varience on codon usage at start and stop site showed variation in codon selection in these sites. This study provides evidence that the dicot genes were subjected to compositional selection pressure.

Keywords. codon; dinucleotide; selection; mutation; genome.

Introduction

Proteins are the building blocks of life, encoded by one or more polypeptide chains. Amino acids are the monomeric units of proteins coded by triplet nucleotides called codons. Except methionine (ATG) and tryptophan (TGG) all the other amino acids are encoded by more than one codon because they have alternative codons, known as synonymous codons. Synonymous codons are not used randomly; some are used most frequently than others. Nonrandom usage of codons from degenerate codon families has been observed because the standard genetic code is not used in its built-in redundancy in the same way by all the species from prokaryotes to eukaryotes. Some organisms (*Candida albicans* and related species) use the nonstandard genetic code to encode proteins, which affect the translational process and codon usage. Codon usage bias (CUB) is a well-studied phenomenon across a wide range of organisms, essentially depicts the unequal usage of synonymous codons in genes. Codon bias offers the opportunity to change the efficiency and accuracy of protein production (Cristina *et al.* 2016).

Previous reports on codon bias study suggest that it is a complex process associated with several factors like natural selection (Ikemura 1981), mutation pressure (Sueoka 1988), genetic drift (Doherty and McInerney 2013), protein structure (Adzhubei *et al.* 1996) gene expression level (Hambuch and Parsch 2005), gene length (Duret and Mouchiroud 1999), GC content (Hu *et al.* 2007), environmental stress (Goodarzi *et al.* 2008), population size (Berg 1996), evolutionary age of genes (Prat *et al.* 2009) etc. During evolution, these factors have forced the genome to adapt its CUB according to the control gene expression and protein production (Gustafsson *et al.* 2004). Moreover, the genes from the same genome show variation in their codon usage pattern (Zhao *et al.* 2016). Hence, genomewide comparison of codon bias patterns between closely related species can help unravel the diverse pat-

Electronic supplementary material: The online version of this article (https://doi.org/10.1007/s12041-018-0892-1) contains supplementary material, which is available to authorized users.

terns (distribution of nucleotides, distribution of genes and composition of genes) and the forces that shape their evolution (Behura and Severson 2012; Subramanian and Sarkar 2015).

In recent years, the rapid increase in genome sequencing and the advancement of sequence-based studies have enabled genomewide alignments between related genomes. A growing body of research indicates that microsynteny is common among dicot genomes (Yan *et al.* 2003). Comparative genetic mapping has demonstrated that the dicots namely *Glycine max*, *Arabidopsis thaliana* and *Medicago truncatula* show high level of synteny (conservation of gene content and order between species) (Yan *et al.* 2003; Zhu *et al.* 2003). In the present study, we have chosen the coding sequences of *G. max*, *A. thaliana* and *M. truncatula* for synonymous codon usage and codon context analysis along with comparative study of codon usage.

Materials and methods

The data used in the present study includes the coding sequences from three dicot genomes, i.e. *G. max (Glycine_max_v2.0), A. thaliana (TAIR10) and M. truncatula (MedtrA17_4.0)* and were retrieved from NCBI nucleotide database (http://www.ncbi.nlm.nih.gov). To minimize sampling errors (Wright 1990), we removed the coding sequences that did not have correct initial and termination codons, that had more than one internal stop codon, and those which are not perfect multiple of three nucleotides. Coding sequences, 2000 from *G. max*, 2500 *A. thaliana* and 2442 *M. truncatula* were shortlisted for the final analysis. A Perl script was developed by the corresponding author to count the individual nucleotide (A, T, G and C) and nucleotide at third codon position (A3, T3, C3 and G3) for each of the coding sequences. The individual count was used to sum up the AT and GC content for each coding sequence. Parity rule 2 (PR2) plots were drawn based on AT-bias [A3/(A3 + T3)] and GC-bias [G3/(G3 + C3)]. GC content at first, second and third codon positions (GC1, GC2 and GC3) were also calculated. GC12 is the average of GC1 and GC2. GC3s value is the frequency of G + C at the third synonymously variable codon position.

Codon usage bias measurement index

Effective number of codons (Nc) is a widely used measure of codon usage bias, affected only by GC3 as a consequence of mutation pressure and genetic drift (Mirsafian *et al.* 2014). Nc quantifies the extent to which the usage of a gene departs from the equal usage of synonymous codons (Wright 1990). It is a measure independent of gene length and composition. Nc value ranges from 20 to 60. Higher Nc indicates lower codon usage bias and lower Nc reveals higher codon usage bias. Nc value in the range 20–45 is generally considered as high codon bias. The Nc value is calculated according to Wright (1990)

$$\mathrm{Nc} = 2 + \frac{9}{F_2} + \frac{1}{F_3} + \frac{5}{F_4} + \frac{3}{F_6},$$

where F_i denotes the average homozygosity for the class with i synonymous codons. The coefficients 9, 1, 5, and 3 come from the number of amino acids belonging to the different classes with degeneracy levels 2, 3, 4 and 6.

Relative synonymous codon usage (RSCU) is another well known measure of codon bias, used to examine the frequency of each synonymous codon that encodes the same amino acid. It is calculated as the ratio of the observed frequency of a codon to the expected frequency (codons are used randomly) (Sharp and Li 1986). RSCU value close to 1 indicates that all the synonymous codons encoding the same amino acid are used equally (Gupta and Ghosh 2001). The index is calculated as follows:

$$\mathrm{RSCU} = \frac{g_{ab}}{\sum_b^{n_a} g_{ab}} n_a,$$

where g_{ab} is the observed number of the ath codon for the bth amino acid which has n_a kinds of synonymous codons.

The codon adaptation index (CAI) is another effective measure of codon usage bias proposed by Sharp and Li (1987). CAI is a measurement of relative adaptedness of the codon usage of a gene towards the codon usage of the highly expressed genes. In other words, it measures the extent of bias towards codons that are known to be preferred in highly expressed genes (Guo *et al.* 2007). CAI values range from 0 to 1. A value of 1 indicates strong codon bias in which the preferred codon is always used, and vice versa (Yang *et al.* 2014). The CAI is calculated as follows:

$$\mathrm{CAI} = \exp \frac{1}{L} \sum \ln Wc(k),$$

where L is the number of codons in the gene and $Wc(k)$ is the w (weight) value for the kth codon in the gene. For each amino acid, the weight of each of its codon in CAI is computed as the ratio between the observed frequency of the codon and the frequency of the most frequent synonymous codon for that amino acid. The CAI defines the frequent codons in highly expressed genes as the translationally preferred codons.

Dinucleotide odds ratio

The odds ratio calculation is commonly used to evaluate dinucleotides, pairs of nucleotides, in coding sequences (Plotkin *et al.* 2004). Odds ratio is the likelihood of observing a dinucleotide in a sequence and is calculated as follows:

$$P_{xy} = \frac{f_{xy}}{f_x f_y},$$

Table 1. Comparison of base composition among the dicot genes.

	No. of sequences	A%	T%	G%	C%	GC%	GC1%	GC2%	GC3%
G. max	2000	27.8	25.9	24.2	21.8	46.0	51.5	40.5	46.1
A. thaliana	2500	28.4	26.7	24.0	20.7	44.6	50.3	40.5	43.1
M. truncatula	2442	29.9	28.9	22.3	18.7	40.9	46.6	38.5	37.6

where, x and y stand for the nucleotides that form dinucleotide xy; and f_x, f_y, f_{xy} denote the frequencies of nucleotide x, nucleotide y, and dinucleotide xy, respectively. Karlin *et al.* (1998) showed that dinucleotides with an odds ratio falling outside the range (0.78–1.23) could be considered as being a more represented or over-represented dinucleotide than normal (Karlin *et al.* 1998).

Correspondence analysis

The correspondence analysis (COA) is a widely used multivariate statistical method used to study the major trends in codon usage variation (RSCU score) (Gupta *et al.* 2004). In COA, all genes are plotted in a 59-dimensional hyperspace according to the RSCU score of 59 sense codons. Sequences in which a given codon is used in a similar fashion lie close to each other on the graph. COA provides a major trend of factors related to codon usage in different gene sets. The major trends in codon usage variation can be determined with relative inertia, according to which the genes are located to investigate the major factors affecting the codon usage pattern.

Software

Codon usage bias measurements were calculated using the program CodonW 1.4.4 (http://codonw.sourceforge.net/) and an in-house Perl script developed by the corresponding author. The 'CodonO' software was used to calculate the synonymous codon usage order index of each gene (Angellotti *et al.* 2007). All the statistical analyses were done using the SPSS software.

Results

Base composition analysis

In the present study on codon usage bias, we have selected genes from three dicots, namely *G. max*, *A. thaliana* and *M. truncatula*. The aforementioned plant genomes show colinearity/high level of synteny among themselves (Yan *et al.* 2004; Mudge *et al.* 2005). To identify and understand the potential impact of nucleotide composition on colinearity between the genomes, the coding sequences from three dicots were investigated (table 1). In support of the previous research on plant genes, here also the genes show bias for the use of A/T within the coding region (Liu and Xue 2005). The nucleobase A showed the utmost use followed by T. The usage frequency of A/T was found maximum in *M. truncatula* followed by *A. thaliana* and the smallest in *G. max*, respectively, whereas, G/C showed the reverse usage pattern of A/T. Genomic GC varies significantly among different species within the same order due to the differences in mutational pressure operating on them (Behura and Severson 2012). The highest global GC content was observed in *G. max* followed by *A. thaliana* and the lowest in *M. truncatula*. The mean ± SD of GC is 46.05 ± 4.58, 44.67 ± 3.40 and 40.95 ± 4.72 for the genes of *G. Max*, *A. thaliana* and *M. truncatula*, respectively. The analysis of GC content at 1st, 2nd and 3rd codon positions showed variation in their usage pattern across the dicots (figure 1 in electronic supplementary material at http://www.ias.ac.in/jgenet/). Differences in GC content were largest at the 3rd codon position (ranges from 25.6% to 85.3%), followed by 2nd codon position (ranged from 12.2% to 82.6%) and 1st codon position (ranged from 17.7% to 77.4%) for the genes of *G. Max*, *A. thaliana* and *M. truncatula*, respectively. Extensive research on codon bias suggests that GC3 is the most important factor for genome evolution, and it also influences the gene expression level (Bellgard *et al.* 2001; Kawabe and Miyashita 2003). In our study, the usage pattern of GC3 directly mimics the genomic GC, i.e. higher the overall GC content bias, higher is the local GC composition (GC3) (figure 2 in in electronic supplementary material).

Codon usage bias analysis

All the synonymous codons are not used uniformly; some are used more frequently than its synonymous partners, influenced mainly by the compositional mutation bias and natural selection (Xiang *et al.* 2015). The relative contribution of mutation and/or selection pressure to the observed codon bias varies from genome to genome (Singer and Hickey 2003). RSCU score was analysed for every codon for each of the three dicot genes (figure 1). For all the amino acids, it was observed that the codons with A/T at 3rd codon position showed the maximum usage frequency among their synonymous partners, and this suggests that genes of these dicots are highly dominated by codons ending with A/T. These results indicate that the codon usage pattern in these dicot species is mostly contributed by compositional constraints. G-ending codon TTG (leucine)

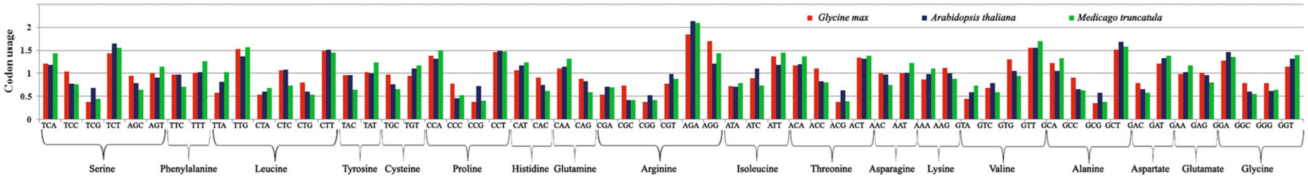
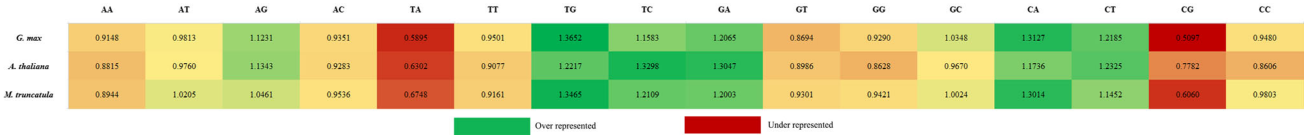


Figure 1. Relative synonymous codon bias analysis of 59 codons.



	AA	AT	AG	AC	TA	TT	TG	TC	GA	GT	GG	GC	CA	CT	CG	CC
G. max	0.9148	0.9813	1.1231	0.9351	0.5895	0.9501	1.3652	1.1583	1.2065	0.8694	0.9290	1.0348	1.3127	1.2185	0.5097	0.9480
A. thaliana	0.8815	0.9760	1.1343	0.9283	0.6302	0.9077	1.2217	1.3298	1.3047	0.8986	0.8628	0.9670	1.1736	1.2325	0.7782	0.8606
M. truncatula	0.8944	1.0205	1.0461	0.9536	0.6748	0.9161	1.3465	1.2109	1.2003	0.9301	0.9421	1.0024	1.3014	1.1452	0.6060	0.9803

🟩 Over represented 🟥 Under represented

Figure 2. Heat map showing distribution of relative abundance of dinucleotides in three plant genomes. Green, red and light yellow indicate over-representative, under-representative and normally used dinucleotides, respectively.

showed the highest usage in *G. max* and *M. truncatula*. For *A. thaliana*, CTT showed utmost usage among all the synonymous codons to encode leucine. Further, for the amino acid lysine, AAG showed almost similar usage frequency to its synonymous partner AAA (favoured codon). The exceptional RSCU score of these two G-ending codons may be due to the presence of TpT and ApA dinucleotides. TCG (S), CCC and CCG (P), CGC and CGG (R), ACG (T) and GCG (A) are the codons that showed the lowest average RSCU. All these rare codons were found to be G/C-ending and also rich in G/C mononucleotides. Favoured codon analysis reveals that the codons with A/T are more preferred compared to G/C-ending codons (figure 1). From our analysis it was observed that four of 18 preferred codons had RSCU values ≥1.6. The over-represented codons are TCT (S), AGA (R), GTT (V), and GCT (A). Except these four codons other preferred codons had RSCU values between 0.60 to 1.59. The under-represented codons, i.e. those with RSCU<0.6 are TCG (S), CCC (P), CCG (P), CGC (R), CGG (R), ACG (T), and GCG (A). The nucleotide composition and RSCU analyses suggest that the selection of the preferred codon is mainly influenced by the compositional properties of the genome, supporting the result of Butt *et al.* (2014). The high tendency to use A/T over G/C mononucleotide-containing codons in AT-rich dicot genome suggests that nucleotide composition, and not the mutation bias, is an important factor of codon usage bias.

We further analysed the synonymous codon usage order (SCUO) of genes and found that *G. Max* had the least average (0.133 ± 0.05); *A. thaliana* showed the average value 0.134 ± 0.07; *M. trucatula* showed the highest average value 0.224±0.10 among the three species. The SCUO analysis showed that a majority of the dicot genes were associated with low codon usage bias. Out of all the genes analysed in the present study, a total of 37 genes showed SCUO > 0.5 (two for *G. Max*, three for *A. thaliana*, and 33 for *M. truncatua*). The orthologous genes of the three genomes were retrieved from EnsemblPlants using BioMart (http://plants.ensembl.org/index.html). A total of 6000 orthologous genes (2000 for each genome) were selected for SCUO analysis. In support of the genomewide genes, the orthologous genes from *M. trucatula* showed the highest average (0.148 ± 0.57). Moreover, it was also found that the codon usage orders of the orthologous genes showed a correlation of 0.562 with the genomewide average SCUOs of the species.

Dinucleotide bias

Dinucleotide usage bias influences gene expression through alternations in translation efficiency. The relative abundance (odds ratio) of the 16 possible dinucleotides was analysed as shown in figure 2, through which several distinct patterns can be observed. The dinucleotide TpG, TpC, GpA, CpA and CpT showed over-representation, whereas, CpG and TpA dinucleotides showed under-representation in all the three genomes. The relative abundance value for each of the over representative and under representative dinucleotides are given in the figure 2. To identify the effect of dinucleotide usage on the codon usage bias, we compared the over-representative and under-representative dinucleotides with preferred and rare codons, respectively. TpG dinucleotides-containing codons are TTG (L), CTG (L), TGC (C), TGT (C) and GTG (V). TTG codon was found to be a preferred codon, whereas CTG was a rare codon based on their RSCU score. For the two-fold degenerative amino acid cysteine, both the synonymous codon contains TpG dinucleotide. GTG showed an RSCU score close to 1. TpC dinucleotide-containing codons are TCA (S), TTC (S), TCG (S), TCT (S), TTC (F), CTC (L), ATC (I) and GTC (V). Of the six synonymous codon for the amino acid serine, four codon contained TpC dinucleotide. Among them, TCA and TCT codons were preferred; TTC and TCG are rare codons that encode the particular amino acid. TTC and CTC showed RSCU score close to 1 (except *M. truncatula*).

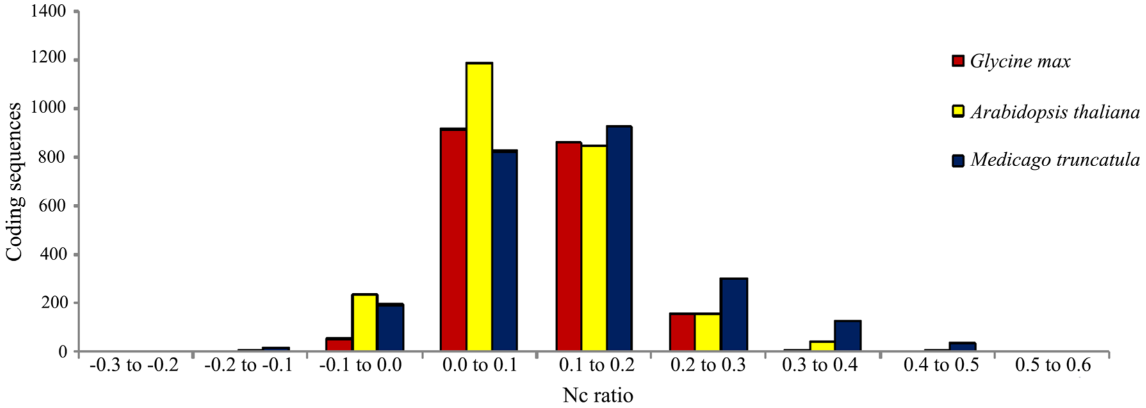


Figure 3. Frequency distribution of (NcExp−NcObs)/NcExp, i.e. Nc ratio for the coding sequences of three dicots.

ATC and GTC are the rare codons that contain the over-representative dinucleotide TpC. GpA dinucleotide-containing codons are CGA (R), AGA (R), GAC (D), GAT (D), GAA(E), GAG (E) and GGA (G). AGA is the preferred codon, whereas, CGA is the rare codon to encode arginine. For the two 2-fold degenerative amino acids aspartate and glutamate, all synonymous codons contain GpA dinucleotide. CpA dinucleotide-containing codons are TCA (S), CCA (P), CAT (H), CAC (H), CAA (Q), CAG (Q), ACA (T) and GCA (A). All the codons encoding histidine and glutamine contain CpA dinucleotide. Except these two amino acids all other synonymous codons with CpA dinucleotide showed RSCU $\geq$ 1. CpT dinucleotide-containing codons are TCT (S), CTA (L), CTC (L), CTG (L), CTT (L), CCT (P), ACT (T) and GCT (A). In the case of leucine, where four of six synonymous codons contain CpT dinucleotide, the codon CTT was found as the preferred one to encode the amino acid. Except leucine, all other codons with CpT dinucleotide showed RSCU score close to 1.5 (preferred codon). The codons with under-representative dinucleotide CpG are TCG (S), CCG (P), CGA (R), CGC (R), CGT (R), CGG (R), ACG (T) and GCG (A), whereas, TpA-containing codons are TTA (L), CTA (L), TAC (Y), TAT (Y) and ATA (I). All the codons with under-representative dinucleotides (except tyrosine where both the codons contain TpA) showed very low RSCU or rare codon usage. Earlier studies on dinucleotide bias analysis suggest that the under-representation of CpG dinucleotide is attributed to mutation that occurs during the methylation process or to selection pressure against methylation (De Amicis and Marchetti 2000; Morton *et al.* (2006)). A few workers have also opined that the underrepresentation of TpA dinucleotides and codons with TpA dinucleotide in plant genomes is related to mRNA stability (Kariin and Burge 1995). From our analysis it could be hypothesised that the increased usage of some dinucleotides and the suppression of some other dinucleotides in coding sequences result in preference of favoured codons and avoidance of rare codons, respectively. This comparative analysis indicates that the composition bias as a factor of dinucleotide usage bias influences the preference of codon usage within a genome.

Effective number of codons

Codon usage bias within the genes was measured using the effective codon usage statistic, Nc (Wright 1990). A lower Nc value indicates high bias. The average Nc was 53.63 and ranged from 44 to 60.31 for *A. thaliana* genes. *G. Max* genes showed the average Nc of 51.33 and ranged from 33.95 to 60.25, whereas, *M. truncatula* showed an average Nc of 48.20 and ranged from 27.33 to 61. Therefore, these genes were not much biased with regard to codon usage, although some genes from *G. max* and *M. truncatula* showed Nc < 35 (high codon bias) indicating the presence of codon usage variation in these genes (Li *et al.* 2016). To measure the difference between the observed codon bias (as analysed by Nc) and expected codon bias when there is no bias in selection, we used a plot of the frequency distribution of (NcExp−NcObs)/NcExp (figure 3).

From figure 3, it is evident that majority of the genes appear in the 0.00–0.10 range, suggesting that most genes show codon bias smaller than the expected codon bias based on random GC3 distribution. Further, a good number of genes in the (NcExp−NcObs)/NcExp plot showed their presence within 0.10–0.20 regions. The value close to zero suggests that mutation pressure (random GC3 distribution) plays a significant role in the formation of codon usage bias. Many genes occupied a position beyond the 0.10 region in the Nc ratio plot. Altogether, these results suggest that mutation pressure might be a factor of codon usage bias, whereas other factors, like selection pressure strongly affected the codon usage bias of the genes in dicots.

Evidence regarding the effect of mutation pressure and natural selection on the evolution of codon usage bias has been put forward based on the findings of several genomes (Sharp *et al.* 2010; Nasrullah *et al.* 2015). To find the effect of these evolutionary forces on codon composition bias,

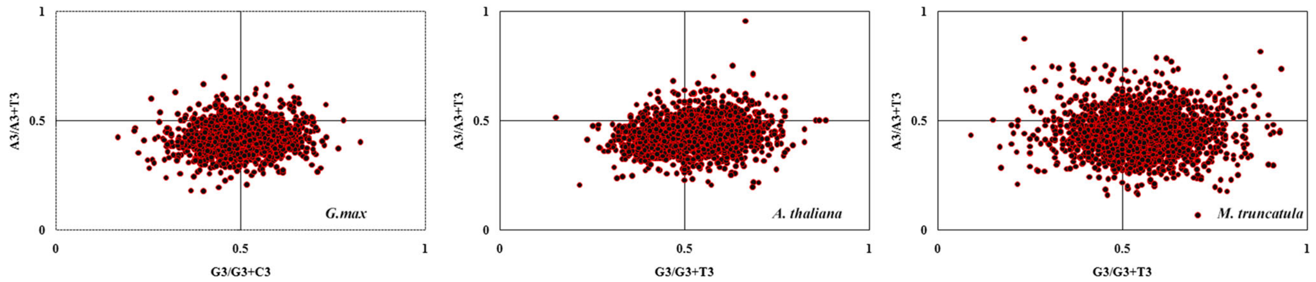


Figure 4. PR2-bias plots of dicot genomes. Genes are plotted based on their GC bias (G3/(G3 + C3)) and AT bias (A3/(A3 + T3)) in the third codon position.

we analysed the correlation between GC_{12s} and GC_{3s}. Three genomes showed very low correlation between average GC content of 1st and 2nd codon positions with that of the 3rd codon position (*G. max*: $r = 0.23, P < 0.01$; *A. thaliana*: $r = -0.01$; *M. truncatula*: $r = 0.25, P < 0.01$).

From the neutrality plot analysis between GC12 and GC3 (figure 3 in electronic supplementary material), it was observed that the slope of the regression line is 0.083, 0.005, and 0.185 for the coding sequences of *G. max*, *A. thaliana* and *M. truncatla*, respectively. This suggests that the effect of mutation pressure is 8.3%, 0.5% and 18.5% while the influence of other factors, for example natural selection, is 91.7%, 99.5% and 81.5% for the coding sequences of *G. max*, *A. thaliana* and *M. truncatla*, respectively (Chen 2013). These results suggest that natural selection played a major role, while mutation pressure played a minor role in determining the pattern of nucleotide usage at different codon positions for the dicot genes (Jia *et al.* 2015).

PR2 bias analysis

From the initial synonymous codon usage bias analysis, it was observed that the dicot genomes prefer A/T-ending codons over G/C-ending synonymous codons. To evaluate more precisely the bias pattern of nucleotides at the wobble position, we performed PR2 analysis (figure 4). In the PR2 plot, if the mean GC bias (G3/(G3 + C3)) equals AT bias (A3/(A3 + T3)) then it suggests that mutation pressure operates on the formation of codon usage pattern of the genome (Zhang *et al.* 2007). Natural selection for codon selection would not necessarily cause proportional use of G and C (A and T). In the PR2 plot, the mean GC-bias of *G. max*, *A. thaliana* and *M. truncatula* was 0.51, 0.53, and 0.55; while their mean AT-bias is 0.43, 0.44, and 0.44, respectively. Here, it was observed that the codons are G and T biased at the wobble position for the genes of *G. max*, *A. thaliana* and *M. Truncatula*. These results suggest that natural selection pressure might play a major role in codon usage bias, and that mutation pressure could be a minor factor.

Table 2. Percentages of three major axes contributing to the codon usage variation among the genes. The low value of axis 1 of *A. thaliana* and *M. truncatula* might be due to the extreme genomic composition.

Axis	*G. max*	*A. thaliana*	*M. truncatula*
1	21.47%	8.96%	6.59%
2	6.16%	5.63%	4.20%
3	4.44%	4.40%	4.10%

Correspondence analysis

Correspondence analysis is widely used to measure the effect of nucleotide composition in codon usage pattern. Further to investigate the effect of composition bias, we performed correspondence analysis on the RSCU values. Correspondence analysis of dicot genes revealed that there is a little variation among the first three axes with respect to their efficiency in explaining the codon usage variation among the genes (except *G. max;* see table 2). The position of each gene on the plane defined by the first three axes is displayed in figure 5, and the distance in the figure represents the variation of RSCU values of different genes. Further, AT-ending codons were labelled with red colour and GC-ending codons with black colour to investigate the effect of nucleotide bias at wobble position to the overall codon usage pattern. The A/T-ending codons were found to be more concentrated, indicating similar codon bias pattern; whereas G/C-ending codons scattered distantly over the plot area. Consistent to RSCU analysis, the T/A-ending codons were used most frequently than G/C-ending codons. It is also obvious from figure 5 that the majority of the codons clustered together around the origin of axes indicating more or less similar codon usage bias. All the rare codons were located distantly from other codons revealing the difference in their usage frequency.

Effect of other sequence-based factors associated with codon usage bias

CAI is known to show a significant correlation with protein abundance in many organisms (Sharp and Li 1987).

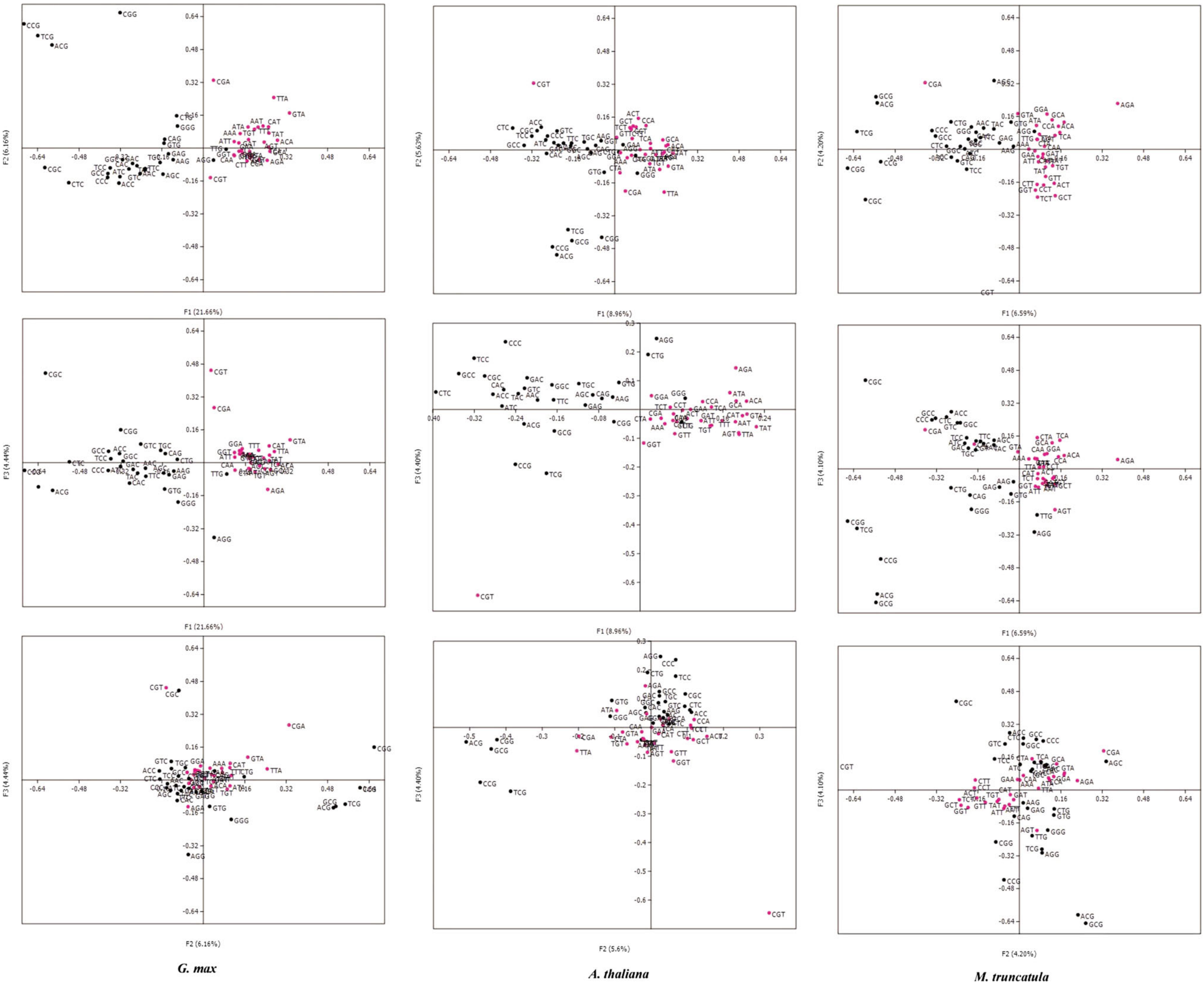


Figure 5. Positions of genes along the three major axes of variation in the correspondence analysis based on RSCU values.

The CAI value ranged from 0.12 to 0.40 with an average of 0.21 for *G. max*; 0.009 to 0.396 with an average of 0.178 for *A. thaliana*; and 0.058 to 0.369 with an average of 0.194 for the *M. truncatula* genes. Correlation coefficients between the nucleotide compositions at third codon position, CAI, aromo, Gravy with codon usage bias are provided in figure 4 in electronic supplementary material. Gravy and GC3 showed significant correlation with gene length (except in *M. truncatula*, see table 1 in electronic supplementary material. Nc showed significant correlation with nucleotide compositions, GC and GC3s. Aromo showed significant correlation with axis 1 at $P < 0.01$ (except *M. truncatula*). In case of *A. thaliana* Gravy showed significant correlation with axis 1 at $P < 0.01$. Nc showed low correlation with CAI. These findings suggest that different sequence-based factors are associated with the codon bias pattern in the dicots.

Position-wise nucleotide bias and codon bias analysis

GC content is not used uniformly in the coding region. Pattern of GC usage shows close association with several other biological factors (e.g. recombination rates) (Ressayre *et al.* 2015). To identify the existence of GC content pattern variation within the genes of plant genomes, we selected the initial 13 codons (39 nucleotides excluding start codon ATG) from the start site and terminal 13 codons (39 nucleotides excluding stop codon) for each coding sequences (cds) and compared the nucleotide composition among them. In the start site, the usage of GC is maximum as compared to the stop site. The global GC content (calculated on the complete cds) varies between the start and stop sites. The graphical representation of GC pattern in three different cds positions (start site, globally and stop site) across the three plant genomes are given in electronic supplementary material figure 7 in. From our

Table 3. Comparison of base composition at the start and stop sites among the dicot genes.

	G. max		*A. thaliana*		*M. truncatula*	
	Start site	Stop site	Start site	Stop site	Start site	Stop site
A	9.73 ± 3.60	11.07 ± 3.26	10.69 ± 3.75	11.26 ± 3.48	11.70 ± 3.69	11.82 ± 3.63
T	10.45 ± 3.90	10.46 ± 3.18	10.38 ± 3.78	10.44 ± 3.49	11.30 ± 4.07	11.43 ± 3.75
G	9.43 ± 3.81	8.86 ± 3.05	9.41 ± 3.44	8.85 ± 3.02	8.29 ± 3.43	8.43 ± 3.15
C	9.37 ± 3.95	8.38 ± 3.08	8.51 ± 3.44	8.42 ± 3.07	7.69 ± 3.45	7.30 ± 3.04

analysis it was observed that in all the three genomes the GC content showed a 5′–3′ decreasing gradient. Previous studies have also found the same pattern of GC usage in several other plant genomes (Serres-Giardi *et al.* 2012, Clement *et al.* 2014). The average usage of A, T, G and C for three plant genomes at start and stop sites is given in table 3. From the comparative composition analysis it was observed that despite the difference in their overall GC content, the genes from all the genomes at the start and stop sites exhibit a common trend in nucleotide usage. Except *G. max*, the genes of other two dicot species showed the highest use of A at start and stop sites followed by T.

To identify the relative bias in codon usage, we analysed the RSCU score of codons for each coding sequence at the start and stop sites. Three plant genomes showed almost similar pattern of codon usage. Analysis of vareince (ANOVA) was carried out to address if there was any difference of codon usage at the start and stop sites. ANOVA identified the codons: TCC, TGC, TCT, TGT, CCC, CAT, CAC, CGT, ATA, AAG, GTG and TGG showing significant difference at $P < 0.05$ in their usage frequency at the start and stop sites for the genes of *A. thaliana*. The codons TCT, TAC, TGC, CTT, CAC, CAG, AGA, AGG, ATC, ACT, AAG, GCT, GAG, GGC in *G. max*, and the codons TCG, TCT, CTC, CCG, AAC, AAA, GCA and GAG in *M. truncatula* showed significant difference at $P < 0.05$. To further study the different properties of start and stop regions, we have performed codon pair bias analysis for the start codon and three stop codons. From our analysis, it was observed that some specific codon pairs are more frequently used as a context of stop codon in these plant genomes. Among all the codons, the A-heading codons showed high usage frequency as a context of stop codon above the average. For *G. max* the G-heading codons as a context of TAA showed the highest usage as compared to A-heading codons. The G-heading codons as a context of TAA in other genomes appeared as the second most preferred codon. The codons AAT and AAA are the most frequent 5 prime context for all the standard stop codons (figure 6 in electronic supplementary material). On the other hand, GCT, GGT, GCG, GAG, GGA, GAA and GAT are the most frequent 3 prime context of the start codon (figure 5 in electronic supplementary material). Behura and Severson (2010) also observed similar nonrandom codon usage pattern in Dipteran and Hymenopteran genomes (Behura and Severson 2012). Start codon pair bias analysis revealed the avoidance of C-heading codons as a context of ATG. Therefore, from the codon context analysis of start and stop codons, it was observed that the genes show a nonrandom usage of codons in the adjacent position of start and stop codons.

Discussion

Indices of codon usage bias were used to measure differences in the occurrence of codon usage that reflect the evolutionary patterns of their genome. Here, the compositional properties and codon bias were estimated for the genes from three dicots, namely *G. max*, *A. thaliana* and *M. truncatula*. The results of composition analysis clearly revealed that the three dicots followed almost similar pattern of nucleotide composition, i.e. genes are A/T biased. The AT-biased genomic architecture of coding sequences may be due to its relationship with evolutionary fitness. Moreover, the synthesis cost of AT is lower than GC nucleotides (Rocha and Danchin 2002). Previous studies on several genomes suggest that genomic GC, in particular, the GC content at third codon position plays a crucial role in gene expression (Kudla *et al.* 2006; Tatarinova *et al.* 2013). Here, the genes showed large variation of GC3s. The variation of GC3s may be due to the differences in gene length and function. Xia *et al.* (2003) showed positive correlation between GC content and gene length. Mutations at GC3 primarily lead to synonymous substitutions, and the selective pressures affecting its composition at 3rd position are different from those acting on the first two codon positions, making it an important tool to study evolution. The genes showed significant correlation between overall GC and GC content at third codon position. This suggests that selection played a crucial role during the evolution of these dicots. Due to selection on codon usage, preferred or favoured codons are used frequently over the coding region in highly expressed genes (Serres-Giardi *et al.* 2012). The preference of favoured codon is driven by selection pressure, whereas the existence of rare codons is due to the action of mutation and genetic drift (Bulmer 1991). In three dicots, RSCU analysis revealed the preference and avoidance of A/T-ending and G/C-ending codons, respectively. Therefore, our results confirm that the dicot genes experience stronger translational selection of A/T-ending codons due to their high genomic AT content. To detect the effect of the occurrences of dinucleotide in codon usage bias, we first calculated the occurrence of 16 possible dinucleotides. From the odds ratio analysis it was observed that no nucleotide was present at the expected frequency. The dinucleotides TpG, TpC, GpA, CpA and CpT showed higher usage frequency than the expected value, whereas, TpA and CpG showed usage less than the expected value. In these dicots, elevated levels of TpG were found along with CpG depletion, which would result from a cytosine-to-thymidine mutation. The under-representation of CpG or TpA dinucleotides has been reported in other plant genomes (De Amicis and Marchetti 2000). These over-represented and under-represented dinucleotide-containing codons also showed RSCU greater and less than 1, respectively. Therefore, it could be hypothesized that the dicot genes follow a pattern of composition bias, which also complements to codon usage bias.

Nc analysis revealed that the genes are partially biased with respect to codon usage, although some genes from *G. max* and *M. truncatula* showed Nc $<$ 35. The effect of mutation and selection pressure on codon usage variation is central to evolutionary studies. Wright (1990) proposed that comparison of the actual distribution of GC3s with the expected distribution under no selection could be indicative, if the codon usage bias of the genes had some influence other than the genomic GC composition. To do so, we used a plot of the frequency distribution of (NcExp−NcObs)/NcExp. The plot revealed that many genes showed observed value close to expected value. Further, a few genes show value within 0.10–0.20, suggesting that, in addition to mutation pressure, some other factors affect the codon bias of the dicots. Moreover, if mutation pressure is the sole source of codon bias then GC or AT should be used proportionally at the synonymous third codon position, where, selection would not necessarily cause proportional use of G and C (A and T). Sueoka (1999) proposed that PR2 plot could be informative to investigate the impact of selection and mutation pressure on codon usage bias. In the PR2 plot, the centre represents the value 0.5, suggesting no biasness between the influence of mutation and selection rates. The nucleotide T and G showed biased usage at wobble position in all the dicot species. From these findings, we can conclude that selection pressure contributed significantly to the codon usage pattern of dicot genes.

To visualize the distribution of genes (based on RSCU) in high dimensional space, we projected them to a 2-D plane spanned by the first and second, first and third, and second and third principal axes by using the COA method. None of the axes were found to contribute a large variation, except *G. max*, where the first axis explains 21.47% of the total variation. Axis 1 showed significant correlation with overall GC and GC3s, which confirms the GC consequence on codon usage pattern. GC-ending codons were found scattered throughout the plot area in contrast to the more concentrated AT-ending codons. This observation suggests the presence of variation due to GC-ending codon usage among the dicot genes. All the three axes showed significant correlations with gravy and aromaticity indicating the possible influence of hydropathic character of protein and aromatic amino acid composition in codon usage variation across dicot genes. Earlier workers have suggested that some genes across organisms are rich in rare codons in the translation initiator region, which leads to increased translational efficiency (Goodman *et al.* 2013). Here the genes showed decreasing GC gradient from initiator to termination region. Comparative nucleotide analysis at the start and stop regions showed the highest usage of A at both start and stop regions followed by T. This reveals that the nucleotides A and T are used with the highest frequency irrespective of the location within the coding region. Moreover, to identify variation in codon usage at the start and stop regions, we performed ANOVA analysis. ANOVA identified that the codons TCC, TGC, TCT, TGT, TAC, TCG, CCC, CAT, CAC, CGT, CTT, CAG, CCG, ATA, AAG, AGA, AGG, ATC, ACT, AAC, AAA, GTG, GCT, GAG, GGC and GCA show significant difference for usage at $P < 0.05$. $3'$ and $5'$ codon context analyses of the start and stop codons showed that some codons are used most frequently as context to the start and stop codons, respectively. Our analysis revealed the avoidance of the C-heading codons as a context of the start codon. The codons AAT and AAA are the most frequent 3 prime context for all the standard stop codons (1.526 $<$ average). On the other hand, GCT, GGT, GCG, GAG, GGA, GAA and GAT are the most frequent 5 prime context of the start codon (1.639 $<$ average). Therefore, from the result of position-dependent codon usage and codon context analysis it is evident that dicot genes show variation in synonymous codon selection not only in the coding region but also in the context of start and stop codons as a pair.

Acknowledgements

Authors are grateful to the University Grants Commission, New Delhi, India, for providing UGC-BSR Fellowship to carry out this research work. Authors are also grateful to the Assam University, Silchar, Assam, India for providing the research facility.

References

Adzhubei A. A., Adzhubei I. A., Krasheninnikov I. A. and Neidle S. 1996 Non-random usage of 'degenerate' codons is related to protein three-dimensional structure. *FEBS Lett.* **399**, 78–82.

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

Behura S. K. and Severson D. W. 2012 Comparative analysis of codon usage bias and codon context patterns between dipteran and hymenopteran sequenced genomes. *PLoS One* **7**, e43111.

Bellgard M., Schibeci D., Trifonov E. and Gojobori T. 2001 Early detection of G + C differences in bacterial species inferred from the comparative analysis of the two completely sequenced Helicobacter pylori strains. *J. Mol. Evol.* **53**, 465–468.

Berg O. G. 1996 Selection intensity for codon bias and the effective population size of *Escherichia coli*. *Genetics* **142**, 1379–1382.

Bulmer M. 1991 The selection-mutation-drift theory of synonymous codon usage. *Genetics* **129**, 897–907.

Butt A. M., Nasrullah I. and Tong Y. 2014 Genome-wide analysis of codon usage and influencing factors in chikungunya viruses. *PLoS One* **9**, e90905.

Chen Y. 2013 A comparison of synonymous codon usage bias patterns in DNA and RNA virus genomes: quantifying the relative importance of mutational pressure and natural selection. *Biomed. Res. Int*. Article ID 406342 (https://doi.org/10.1155/2013/406342).

Clement Y., Fustier M. A., Nabholz B. and Glemin S. 2014 The bimodal distribution of genic GC content is ancestral to monocot species. *Genome Biol. Evol.* **7**, 336–348.

Cristina J., Fajardo A., Sonora M., Moratorio G. and Musto H. 2016 A detailed comparative analysis of codon usage bias in Zika virus. *Virus Res.* **223**, 147–152.

De Amicis F. and Marchetti S. 2000 Intercodon dinucleotides affect codon choice in plant genes. *Nucleic Acids Res.* **28**, 3339–3345.

Doherty A. and McInerney J. O. 2013 Translational selection frequently overcomes genetic drift in shaping synonymous codon usage patterns in vertebrates. *Mol. Biol. Evol.* **30**, 2263–2267.

Duret L. and Mouchiroud D. 1999 Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. *Proc. Natl. Acad. Sci. USA* **96**, 4482–4487.

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

Goodman D. B., Church G. M. and Kosuri S. 2013 Causes and effects of N-terminal codon bias in bacterial genes. *Science* **342**, 475–479.

Guo X., Bao J. and Fan L. 2007 Evidence of selectively driven codon usage in rice: implications for GC content evolution of Gramineae genes. *FEBS Lett.* **581**, 1015–1021.

Gupta S. K. and Ghosh T. C. 2001 Gene expressivity is the main factor in dictating the codon usage variation among the genes in Pseudomonas aeruginosa. *Gene* **273**, 63–70.

Gupta S. K., Bhattacharyya T. K. and Ghosh T. C. 2004 Synonymous codon usage in Lactococcus lactis: mutational bias versus translational selection. *J. Biomol. Struct. Dyn.* **21**, 527–536.

Gustafsson C., Govindarajan S. and Minshull J. 2004 Codon bias and heterologous protein expression. *Trends Biotechnol.* **22**, 346–353.

Hambuch T. M. and Parsch J. 2005 Patterns of synonymous codon usage in Drosophila melanogaster genes with sex-biased expression. *Genetics* **170**, 1691–1700.

Hu J., Zhao X., Zhang Z. and Yu J. 2007 Compositional dynamics of guanine and cytosine content in prokaryotic genomes. *Res. Microbiol.* **158**, 363–370.

Ikemura T. 1981 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**, 389–409.

Jia X., Liu S., Zheng H., Li B., Qi Q., Wei L. *et al*. 2015 Non-uniqueness of factors constraint on the codon usage in *Bombyx mori*. *BMC Genomics* **16**, 356.

Kariin S. and Burge C. 1995 Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.* **11**, 283–290.

Karlin S., Mrazek J. and Campbell A. M. 1998 Codon usages in different gene classes of the Escherichia coli genome. *Mol. Microbiol.* **29**, 1341–1355.

Kawabe A. and Miyashita N. T. 2003 Patterns of codon usage bias in three dicot and four monocot plant species. *Genes Genet. Sys.* **78**, 343–352.

Kudla G., Lipinski L., Caffin F., Helwak A. and Zylicz M. 2006 High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS Biol.* **4**, e180.

Li X., Song H., Kuang Y., Chen S., Tian P., Li C. *et al.* 2016 Genome-wide analysis of codon usage bias in Epichloe festucae. *Int. J. Mol. Sci.* **17**, 1138.

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

Mirsafian H., Mat Ripen A., Singh A., Teo P. H., Merican A. F. and Mohamad S. B. 2014 A comparative analysis of synonymous codon usage bias pattern in human albumin superfamily. *Sci. World J.* **2014**, 639682.

Morton B. R., Bi I. V., McMullen M. D. and Gaut B. S. 2006 Variation in mutation dynamics across the maize genome as a function of regional and flanking base composition. *Genetics* **172**, 569–577.

Mudge J., Cannon S. B., Kalo P., Oldroyd G. E., Roe B. A., Town C. D. *et al.* 2005 Highly syntenic regions in the genomes of soybean, Medicago truncatula, and Arabidopsis thaliana. *BMC Plant Biol.* **5**, 1.

Nasrullah I., Butt A. M., Tahir S., Idrees M. and Tong Y. 2015 Genomic analysis of codon usage shows influence of mutation pressure, natural selection, and host features on Marburg virus evolution. *BMC Evol. Biol.* **15**, 1.

Plotkin J. B., Dushoff J. and Fraser H. B. 2004 Detecting selection using a single genome sequence of *M. tuberculosis* and *P. falciparum*. *Nature* **428**, 942–945.

Prat Y., Fromer M., Linial N. and Linial M. 2009 Codon usage is associated with the evolutionary age of genes in metazoan genomes. *BMC Evol. Biol.* **9**, 285.

Ressayre A., Glemin S., Montalent P., Serre-Giardi L., Dillmann C. and Joets J. 2015 Introns structure patterns of variation in nucleotide composition in *Arabidopsis thaliana* and rice protein-coding genes. *Genome Biol. Evol.* **7**, 2913–2928.

Rocha E. P. and Danchin A. 2002 Base composition bias might result from competition for metabolic resources. *Trends Genet.* **18**, 291–294.

Serres-Giardi L., Belkhir K., David J. and Glemin S. 2012 Patterns and evolution of nucleotide landscapes in seed plants. *Plant Cell* **24**, 1379–1397.

Sharp P. M. and Li W.-H. 1986 An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol.* **24**, 28–38.

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

Sharp P. M., Emery L. R. and Zeng K. 2010 Forces that influence the evolution of codon bias. *Phil. Trans. R. Soc. London, Ser. B* **365**, 1203–1212.

Singer G. A. and Hickey D. A. 2003 Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene* **317**, 39–47.

Subramanian A. and Sarkar R. R. 2015 Comparison of codon usage bias across Leishmania and Trypanosomatids to understand mRNA secondary structure, relative protein abundance and pathway functions. *Genomics* **106**, 232–241.

Sueoka N. 1988 Directional mutation pressure and neutral molecular evolution. *Proc. Natl. Acad. Sci. USA* **85**, 2653–2657.

Sueoka N. 1999 Translation-coupled violation of parity rule 2 in human genes is not the cause of heterogeneity of the DNA G+C content of third codon position. *Gene* **238**, 53–58.

Tatarinova T., Elhaik E. and Pellegrini M. 2013 Cross-species analysis of genic GC3 content and DNA methylation patterns. *Genome Biol. Evol.* **5**, 1443–1456.

Wright F. 1990 The 'effective number of codons' used in a gene. *Gene* **87**, 23–29.

Xia X., Xie Z. and Li W. H. 2003 Effects of GC content and mutational pressure on the lengths of exons and coding sequences. *J. Mol. Evol.* **56**, 362–370.

Xiang H., Zhang R., Butler III R. R., Liu T., Zhang L., Pombert J.-F. *et al.* 2015 Comparative analysis of codon usage bias patterns in microsporidian genomes. *PLoS One* **10**, e0129223.

Yan H., Mudge J., Kim D., Shoemaker R., Cook D. and Young N. 2004 Comparative physical mapping reveals features of microsynteny between *Glycine max*, *Medicago truncatula*, and *Arabidopsis thaliana*. *Genome* **47**, 141–155.

Yan H. H., Mudge J., Kim D. J., Larsen D., Shoemaker R. C., Cook D. R. *et al.* 2003 Estimates of conserved microsynteny among the genomes of Glycine max, Medicago truncatula and Arabidopsis thaliana. *Theor. Appl. Genet.* **106**, 1256–1265.

Yang X., Luo X. and Cai X. 2014 Analysis of codon usage pattern in Taenia saginata based on a transcriptome dataset. *Parasit. Vectors* **7**, 1–11.

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

Zhao Y., Zheng H., Xu A., Yan D., Jiang Z., Qi Q. *et al.* 2016 Analysis of codon usage bias of envelope glycoprotein genes in nuclear polyhedrosis virus NPV and its relation to evolution. *BMC Genomics* **17**, 677.

Zhu H., Kim D. J., Baek J. M., Choi H. K., Ellis L. C., Kuester H. *et al.* 2003 Syntenic relationships between Medicago truncatula and Arabidopsis reveal extensive divergence of genome organization. *Plant Physiol.* **131**, 1018–1026.

Corresponding editor: UMESH VARSHNEY