Patterns of genome-wide codon usage bias in tobacco, tomato and potato

Ali Mostafa Anwar*,1, Maha Aljabri*bc and Mohamed El-Soda*

aDepartment of Genetics, Faculty of Agriculture, Cairo University, Giza, Egypt; bDepartment of Biology, Faculty of Applied Sciences, Umm Al-Qura University, Makkah, Saudi Arabia; cResearch Laboratories Centre, Faculty of Applied Sciences, Umm Al-Qura University, Makkah, Saudi Arabia

ABSTRACT

Understanding codon usage patterns and their shaping factors in members of the Solanaceae family may reflect important features of their evolutionary history. Here, we investigated codon usage bias (CUB) over the whole genome of tobacco, tomato and potato. Analyzing the effective number of codons (ENc) revealed a weak CUB over the whole genome of the three genomes. The correlation analysis between the ENc and the nucleotide content at the third codon position indicated that genes with high ENc have high A/T and low G/C contents. Analyzing nucleotide composition and the relative synonymous codon usage confirmed that the three genomes are AT-rich and favor codons ending with T and/or A. Comparing CUB in each of the three genomes revealed that approximately 50% and 45% of the genes were dominated by mutational and natural selection, respectively. The remaining 5% of the genes were affected by combined factors of mutation and translation selection. The first four axes of the correspondence analysis (CA) were the major contributors to the variation within the codon sequences (CDS) of the three genomes. Results of a multi-regression analysis between the 4 axes and the CUB indices verified the fact that natural selection and directional mutation for base composition along with other factors could describe the variations in the CUB within the CDS of the three genomes. In conclusion, the CUB of the three species showed the same weak trend and that mutation and natural selection were the main factors influencing the nuclear genes of the three genomes.

Introduction

During the translation process from mRNAs to proteins, genetic information is transmitted in the form of triple nucleotides named codons to encode amino acids. Apart from methionine (Met) and tryptophan (Trp), each amino acid can be encoded by multiple codons which are known as synonymous codons [1]. Several studies on different organisms reported that synonymous codons are not used uniformly within the same species nor between different species, while translating genes into proteins, a phenomenon called ‘synonymous codon usage bias (SCUB) or codon usage bias (CUB) [2, 3]. Hence, each organism has its optimal codons that are used more frequently in highly expressed genes than in lowly expressed genes [4]. Directional mutation and natural selection are the two main factors shaping CUB of an organism. If mutational bias solely shapes the CUB of an organism, the four base content should be observed uniformly within a codon family in a gene [5, 6]. There are other factors that influence CUB of an organism such as nucleotide composition, gene length [7], tRNA abundance [8], codon hydropathy and DNA replication initiation site [9], and expression level [10]. Investigating the CUB of an organism facilitates a better understanding of its molecular evolution, gene expression, host-pathogen co-adaptation and predicted optimal codons. Investigating the CUB of an organism facilitates a better understanding of its molecular evolution, gene expression, host-pathogen co-adaptation and predicted optimal codons. The later facilitates designing highly expressed genes and constructing suitable cloning vectors [11]. Various

CONTACT

Mohamed El-Soda mohamed.elsoda@agr.cu.edu.eg Department of Genetics, Faculty of Agriculture, Cairo University, Giza, 12613, Egypt.

*Present address: Proteomics and Metabolomics Unit, Basic Research Department, Children’s Cancer Hospital, 57357 Cairo, Egypt.

© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
studies have quantified CUB using the position and the total GC content \([12, 13]\). Those studies illustrated that the synonymous third codon position \((GC_3)\), also known to be linked with translation optimization, is a strong predictor of CUB. Different approaches have been used to substantiate the degree of CUB and to compare the relative effects of natural selection and mutation bias in shaping codon usage. The most widely used approaches are, the effective number of codons (ENc), the codon adaptation index (CAI), the relative synonymous codon usage (RSCU) and the translation selection \((P2)\) index \([4]\). The ENc measures the bias of using a smaller subset of codons apart from the equal use of synonymous codons irrespective of gene length. Generally, highly expressed genes should have a high level of codon preference and hence low ENc values. Besides, ENc can be an indicator of codon usage regarding mutational bias \([14]\).

In the case of CAI, a reference set of highly expressed genes (e.g. ribosomal genes) is used to serve as an indicator of gene expression levels and natural selection, where both help in shaping the pattern of codon usage. Also, CAI can be used as an indicator of selection for a bias toward translational efficiency and/or accuracy \([15]\). RSCU calculates the observed count in relation to the expected count of each analyzed codon. P2 index measures the bias of anticodon-codon interactions, from which the translation efficiency can be inferred. Neutrality plot \([16]\) and parity rule 2 (PR2)-bias plot analysis have also been used to examine the influence of natural selection, mutation or both on CUB \([17]\).

Solanaceae family is one of the biggest families within the angiosperms. This family consists of approximately 3,000–4,000 species and contains a number of major economically important species all over the world such as tobacco \((Nicotiana tabacum)\), tomato \((Solanum lycopersicum)\) and potato \((Solanum tuberosum)\).

Not much information is available on the factors influencing CUB in tobacco, tomato and potato. For example, a previous study on the tomato cultivar Micro-Tom using the available non-redundant full-length cDNA (nrFLcDNA) and expressed sequence tags (ESTs), revealed that GC3 content and CAI were the major factors affecting the CUB \([18]\).

To increase our understanding of the CUB and the main factors influencing it in the three genomes, we designed the present study. Our results revealed similar weak CUB over the whole genome of the three species. We also showed that mutation and natural selection on translation efficiency and/or accuracy were the main shaping factors influencing the nuclear genes of the three species.

**Materials and methods**

**Coding sequence data and nucleotide composition analysis**

Coding sequences of \(Nicotiana tabacum\) (tobacco) (RefSeq assembly accession: GCF_000715135.1) were downloaded from the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov). Coding sequences of \(Solanum lycopersicum\) (tomato) (Accession: GCA_000188115.3) and \(Solanum tuberosum\) (potato) (Accession: GCA_000226075.1) were obtained from the Ensembl Plants FTP server (ftp://ftp.ensemblgenomes.org/pub/plants/release-42).

Nucleotide content/frequencies in the third codon position \((A3, T3, G3 and C3)\) were calculated using a self-developed python script. In addition, GC content at different codon positions, i.e. GC1, GC2 and GC3, as well as the overall GC and GC12, were assessed. Averages were computed to examine the association between codon usage variation and base composition.

**Effective number of codons**

The ENc was calculated using the R package called cor-don. The ENc values range from 20–61, with low values indicate high levels of CUB and high values indicate weak or no bias, i.e. more codons being used for each amino acid. To verify the relationship between ENc and GC3s, ENc-plot was drawn where the expected ENc values from GC3s (denoted by ‘S’) were determined according to the following equation \([19]\):

\[
\text{ENC}_{\text{expected}} = 2 + S \frac{29}{S^2 + (1-S)^2}
\]

Using the above equation, plots of the expected fitting curve of ENc values were drawn using ggplot package in R \([20]\), and then ENc values versus GC3s values for each gene were plotted. If the distribution of the plotted genes is along/near the curve, then the codon usage bias is only affected by the mutation. If the distribution of the plotted genes is below the curve, then the codon usage bias is affected by the selection and other factors.

**Codon adaptation index**

The python Package CAI \([21]\) was used to calculate CAI using the following equation \([15]\):

\[
\text{CAI} = \exp \left( \frac{1}{L} \sum_{k=1}^{L} \ln w_{c(k)} \right)
\]
where \( L \) is the count of codons in the gene and \( w_{c(k)} \) is the relative adaptive value for the \( k \)th codon in the gene. CAI values range from 0 to 1 with 0 for genes with even usage of all synonymous codons, and 1 for genes with consistent usage of specific codons.

**Relative synonymous codon usage**

The R package seqinR [22] was used to calculate RSCU using and the following equation [15]:

\[
RSCU = \frac{O_{ac}}{\frac{1}{k_a} \sum_{c \in C_a} O_{ac}}
\]

where \( O_{ac} \) is the count of codon \( c \) for an amino acid and \( k_a \) is the number of synonymous codons.

From the above equation, if the value of RSCU is less than one, then a codon is observed less frequently than the average codon usage. RSCU becomes equal to one when the codon is observed as the average, and more than one when the codon is observed more frequently than the average codon usage [23].

**Translational selection index**

Translational selection index (P2) for each CDS was calculated using python and the following formula:

\[
P2 = \frac{WWC + SSU}{WWY + SSY}
\]

where \( W = A \) or \( U(T) \), \( S = G \) or \( C \), and \( Y = C \) or \( U(T) \). Then the P2-index for the whole CDS for each genome was calculated by taking the average.

P2 values correlate positively with the level of gene expression, i.e. high for highly expressed genes and low for lowly expressed genes [4]. If P2 is more than 0.5, there is a bias toward translation selection in the tested CDS.

**Neutrality plot**

Neutrality plots were drawn using ggplot package in R [20]. The neutrality plot was used to explore the pattern of codon usage. The average GC content at the first and the second codon positions (GC12) was plotted against the GC content at the third codon position (GC3). A regression plot with a slope equal to or near 0 indicates that natural selection influences codon usage, while a slope equal to or near 1 indicates that mutation influences codon usage [16].

**Parity rule 2-plot analysis**

Using nucleotide content at the third codon position (A3, T3, G3, and C3) for each gene, AT-bias (A3/(A3 + T3)) and GC-bias (G3/(G3 + C3)) were estimated and used in drawing the PR plot as the ordinate and the abscissa, respectively, with both coordinates equal to 0.5, where A = T and G = C [24]. If genes are scattered equally over the plot, then the CUB is likely to be solely caused by the mutation [17]. PR plot was drawn using ggplot package in R [20].

**Determination of putative optimal codons**

Five percent of the highest and lowest expressed genes were determined using the ENc values. Then, the RSCU values for each codon are compared between the highest and lowest expressed genes using the chi-square test. The codon that is used significantly more in the highly expressed genes than in the lowly expressed genes is regarded as an optimal codon [25].

**Correspondence analysis**

To examine the variations in the codon usage bias among all the CDS, we performed multivariate statistical analysis on the 59 codons, i.e. the 61 synonymous codons excluding Met and Trp codons. One way to do that was using correspondence analysis (CA) [17] by plotting the group of genes on the continuous axes in multidimensional space according to the trends affecting the synonymous codon usage within the gene groups. Two types of CA were employed on codons to check the factors affecting the choice of amino acids with respect to the codons being used [26]. The first CA used the relative codon usage frequencies and the second CA used the amino acid usage. The correlation between the first 4 axes of all CA types (the major contribution to the codon usage variation) and CAI, ENc and GC3 was estimated. CA analysis was implemented using CodonW software (http://codonw.sourceforge.net/).

**Results**

**Codon base composition**

To facilitate our analysis of the three genomes, and to investigate CUB and ENc, we used 84,255, 33,697 and 56,210 coding sequences (CDS), i.e. the total count of codons in all CDS of the genome (Table 1) of tobacco, tomato and potato, respectively. The total percent of GC observed here was 42.61%, 41.28%
and 42.50%, while the GC3 percent, i.e. percent of GC at third codon position, were 38.48%, 37.01% and 38.22% for tobacco, tomato and potato, respectively (Table 1). Analyzing each nucleotide at the synonymous third codon position (A3s, T3s, C3s, and G3s) in the genomes of the three species, tobacco, tomato and potato, indicates that the mean values of A3s and T3s are significantly higher ($t$-test, $p < 0.01$) than the mean values of G3s and C3s (Figure 1-i). The neutrality correlation analysis revealed significant, $P$-value <0.05, positive correlation between GC12 and GC3, with slope approximately equal to zero, i.e. 0.000311, 0.00105, 0.00102, for tobacco, tomato and potato, respectively.

The synonymous codon usage

To compare the genomes of the three species and to measure the strength of CUB, ENc was calculated. The three species showed similar weak CUB with ENc values of 58.31, 57.87 and 57.67 for tobacco, tomato and potato, respectively. The ENc values for each gene of the three genomes were significantly negatively correlated with their GC, GC12, i.e. the average of GC1 + GC2, and GC3 nucleotides contents (Table 2). The highest negative correlation, i.e. $r = -0.69$, was observed for potato between GC3 and ENc. In the case of tobacco and tomato, GC3 and ENc were significantly negatively correlated, i.e. $r = -0.43$ and $-0.31$, respectively.

In the ENc plot (Figure 1-ii), the plotting values of ENc against GC3s content illustrated the effect of the base contents on CUB in the three CDSs of the studied genomes. Coding sequences (CDSs) were gathered around the expected ENc fitting curve, in case of high ENc values, weak bias. In contrast, CDSs with low ENc values (strong bias), were found to be below the expected curve.

Table 1. Number of coding sequences (CDS), number of codons and percent of GC content at different positions of codons of the three analyzed genomes, i.e. tobacco, tomato and potato.

|         | Tobacco | Tomato | Potato |
|---------|---------|--------|--------|
| Number of CDS | 84,255  | 33,697 | 56,210 |
| Number of codons | 34,958,895 | 12,026,147 | 16,963,135 |
| GC%      | 42.6% (SD = 3.4) | 41.3% (SD = 4.3) | 42.5% (SD = 4.2) |
| GC12%    | 44.7% (SD = 4.0) | 43.4% (SD = 5) | 44.6% (SD = 4.8) |
| GC3%     | 38.5% (SD = 6.1) | 37.0% (SD = 6.95) | 38.2% (SD = 6.9) |

Note: Number of CDS = the total count of codon sequence of a genome. Number of codons = the total count of codons in all CDS of a genome. GC% refers to the total percent of GC, GC12% = percent of GC at first and second positions, GC3% = percent of GC at third codon position. SD refers to standard deviation.

The synonymous codon usage features and the determination of putative optimal codons

In general, RSCU values were greater than 1, meaning a codon is observed more than expected, for 28 codons in each of the three genomes. These high-frequency codons ended with either A or T in 24 codons, where 20 of these codons are common in the three genomes i.e. AAA, AAT, ACA, ACT, AGA, ATT, CAA, CAT, CCA, CCT, CTT, GAA, GAT, GCA, GCT, GGA, GTA, GTC, TGT, TCT, TCA, TAT. Based on the correlation between RSCU and ENc for each gene in each genome, 25 optimal codons ending with either A (7) or T (16) were determined in tobacco, with two additional codons ending with C (Table 3). In tomato, 30 optimal codons ended either with A (13), T (15) or G (2). In potato, 26 optimal codons were found to end either with A (9), T (16) or G(1) (Table 3).

PR2-bias plot

Figure 1(iii) shows the PR2-bias plots of the three genomes, where CDS of the three genomes were distributed on the left quartiles along the ordinate, or the upper left quartile along the abscissa. The numbers of CDS were scattered on the G>C plane more than the G<C plane.

Translational selection (P2) index

Analyzing the P2-index for each gene in the three genomes revealed similar effects of natural selection on translational efficiency and/or accuracy in tobacco and tomato. For example, out of 84,255 and 33,697 total CDS in tobacco and tomato, respectively, 52% showed low selection effect with P2 < 0.5, 4% had moderate selection effect with P2 = 0.5, and 44% had high selection effect with P2 > 0.5. In the case of the potato genome, the result was slightly different, where from the total of 56,210 CDS; 49% of the CDS had low selection effect with P2 < 0.5, 4% had moderate selection effect with P2 = 0.5, and 47% had high selection effect with P2 > 0.5. For the three genomes, it was observed that both SSU and WWU were higher than SSC and WWC (Table 4), where W = A or U, S = C or G. Hence among the two pyrimidines (U and C) the preference tends to be U (T) in the third position of a codon.

Correspondence analysis using codon usage frequencies

The result for correspondence analysis using the relative synonymous codon usage frequencies (CA-RSCU)
for each genome showed that CA axes could explain 32%, 44% and 46%, respectively, in tobacco, tomato and potato. The first four axes explained most of the CUB variation found in the CDS of the three genomes; therefore correlation analysis between the first four axes and the calculated CUB indices (CAI, ENc, and GC3) was performed (Table 5). To examine the variation within the amino acid, CA using the amino acid

Table 2. The correlation analysis between the effective number of codons (ENc) and each of GC, GC12 and GC3 in the three genomes.

|        | Tobacco | Tomato | Potato |
|--------|---------|--------|--------|
| GC     | GC12    | GC3    | GC     | GC12   | GC3    |
| ENc    | −0.32*  | −0.09* | −0.43* | −0.27* | −0.14* |

Note: GC is the total percent of GC content, GC12 refers to the total percent of GC at the first and second positions, and GC3 refers to the percent of GC at the third codon position. Star sign (*) indicates significant result at p-value < 0.05.
Our results agree with the previous studies that investiga-
ted genomes favor using codons ending with T or A more frequently than codons ending with C or G. This finding suggests that there are similar codon base compositions at the genome level of tobacco, tomato and potato. Our results agree with the previous studies on the genomes of dicot species, i.e. cotton [4], tomato [18], Arabidopsis thaliana [27, 28], peas, tobacco [27], where they show that dicot genomes have low GC-content and tend to use A/T ending codons. Earlier studies showed significant over-representation of codons ending with A/T in dicot species, including tomato and potato, more than in monocot species [29, 30]. Neutrality correlation analysis establishes a relationship between GC12 and GC3 contents to show the role of either natural selection or mutation in shaping the CUB. Our neutrality plots displayed a significant correlation between GC12 and GC3 contents in the three genomes suggesting that mutational stress plays a role in shaping the codon usage preferences. On the other hand, and with a regression neutrality plot with a slope almost equal to zero, natural selection can be suggested to play a role in shaping the codon usage preferences.

The negative correlation between the ENc, with values of 58.31, 57.87 and 57.67 for tobacco, tomato and potato, respectively, and CUB indicate low or weak CUB meaning general random codon preference. Also, weak CUB was observed for CDSs with high ENc values, i.e. around the expected ENc curve in Figure 1(ii), indicating that CUB in the three genomes is affected by the mutation. In addition, the significant negative correlations observed between the ENc and each of the GC, GC12 and GC3 indicate that the base content at the first, second and third positions of the synonymous codons directly affects the level of CUB. Moreover, plotting the ENc values for the CDS versus the GC3s of the three genomes indicates that the codon usage pattern of CDS might be formed by the combined impacts of directional mutation and natural selection.

### Table 3. Putative optimal codons determined in each of the three genomes.

| Tobacco | Tomato | Potato |
|---------|--------|--------|
| AAA AAT ACT AGA AGT ATT CAA CAT CCT | AAT ACT ACA AGT ATT CCA | ACA AAT ACT AAA AGT ATT CCA |
| CGA CGC CGT CCT GAA CAT GCT | CCT CCT CAA CTA CAT TTG GCT | CGT CCT CAA CTA CAT TTG GCT |
| GGA GTT TAT TCC TTC TGT TTA TTT | AGA CAT GTT AAA TAT TTA TCT TGT | GGA GTT GAA TAT TTA TCT TGT |

Note: Underlined codons written in bold are common between the three genomes.

### Table 4. P2 index results and the average counts of codons. P2-index = the average of P2-index for all the CDS.

| Tobacco | Tomato | Potato |
|---------|--------|--------|
| P2-index | 0.5 | 0.3 | 0.5 |
| WWC | 22.6 | 20.1 | 16.9 |
| SSU | 30.5 | 26.3 | 22.3 |
| WWU | 42.1 | 37.8 | 30.9 |
| SSC | 12.2 | 9.4 | 8.6 |

Note: W refers to A or U(T), S refers to C or G. C and U refer to the two nucleotides C and U, respectively. WWC = the average count of WWC codons for all the CDS. WWU = the average count of WWU codons for all the CDS. SSC = the average count of SSC codons for all the CDS. SSU = the average count of SSU codons for all the CDS.

### Table 5. The correlation between the correspondence analysis and each of GC3, ENc and CAI.

| | CA-RSCU | | CA-AA | |
|---|---|---|---|---|
| | GC3 | ENc | CAI | GC3 | ENc | CAI |
| Tobacco | Axis 1 | −0.069* | 0.055* | 0.566* | −0.053* | 0.022* | −0.556* |
| | Axis 2 | 0.094* | −0.096* | 0.204* | −0.020* | 0.011* | −0.323* |
| | Axis 3 | −0.118* | 0.097* | −0.160* | 0.017* | −0.002 | 0.217* |
| | Axis 4 | 0.041* | 0.024* | 0.204* | 0.168* | −0.104* | −0.271* |
| Tomato | Axis 1 | −0.751* | 0.370* | −0.170* | 0.009 | 0.027* | 0.241* |
| | Axis 2 | −0.267* | 0.071* | 0.324* | 0.007 | −0.135* | 0.082* |
| | Axis 3 | 0.229* | 0.149* | 0.190* | 0.037* | −0.040* | −0.083* |
| | Axis 4 | 0.092* | 0.087* | 0.091* | −0.077* | 0.009 | −0.140* |
| Potato | Axis 1 | 0.327* | −0.201* | −0.384* | 0.112* | −0.084* | 0.553* |
| | Axis 2 | 0.224* | 0.203* | −0.373* | −0.010* | 0.014* | −0.201* |
| | Axis 3 | 0.206* | −0.157* | 0.132* | 0.232* | −0.144* | −0.020* |
| | Axis 4 | 0.127* | 0.046* | 0.181* | −0.332* | 0.211* | 0.330* |

Note: CA-RSCU = correspondence analysis based on the frequency of relative synonymous codon usage. CA-AA = correspondence analysis based on amino acid usage. GC3 is the GC content at the third codon position. ENc = the effective number of codons. CAI = the codon adaptation index. Star sign (*) indicates significant result at p-value < 0.05.
To determine the putative optimal codons for the three genomes, RSCU analysis was performed and revealed 26 common codons ending with T or and A. Those 26 codons were observed more frequently than expected. This finding supports a high degree of uniformity among the three genomes. A similar preference of optimal codons was recently reported among four different cotton species [4]. Taking advantage of the positive correlation between the CUB and gene expression, putative optimal codons could be predicted [10]. Therefore, we conclude that genes with weak CUB would have lower G3s, and C3s, and higher T3s and A3s values. Altogether, our results indicate that genes of the three investigated genomes prefer to use high expression codons ending with pyrimidines (T/A). A similar trend was reported for the optimal codons in chloroplast genes among 12 Solanum species, where the chloroplast genes were found to prefer A/T ending codons, and 66% of the optimal codons had A/U at the third codon position [25].

Despite the GC content of a gene, using PR2-bias plot facilitates identifying a gene with CUB formed only by nucleotide composition, i.e. having even distribution between G3 and C3, as well as A3 and T3 [31]. In the present study, the PR-2 bias plots of the three genomes showed inequality between A + T and G + C at the third codon positions indicating that mutation, selection and other factors such as gene expression, gene length, etc. were the main factors defining the codon usage patterns. Similar asymmetry between pyrimidines and purines was previously observed in chloroplast genes among 12 Solanum species [25] as well as in four cotton species [4].

Normalized values of translational selection (P2-index) range between 0 and 1, where a gene with P2-index equal to 1 has the largest translation selection. The CDS of the three genomes investigated here showed an average P2-index equal to 0.5, indicating that the CDS of the three genomes, in general, are dominated by mutation pressure rather than translation selection. Further analysis showed that 44%, 44% and 47% of the genes in the genome of tobacco, tomato and potato, respectively, have a P2-index more than 0.5, indicating that translational selection played the dominant role over mutation pressure in the codon usage. Altogether, our results illustrated that not only one factor, but both translation and mutation affect the CUB of the three genomes.

To verify the factors affecting the codon usage in the three genomes, correlation analysis was carried out between CAI, ENc and GC3 and the first 4 axes of CA. The results showed that the first 4 axes were the major contributors to the codon usage variation for the three genomes. In tobacco, a significant correlation was found between axis-1 of CA-RSCU and CAI ($r = -0.57$, $p$-value < 0.05) which may support the hypothesis that selection for translation efficiency and/or accuracy explains most of the variation of codon usage. Furthermore, CA using amino acid usage (CA-AA) showed strong correlation between axis-1 and CAI ($r = -0.56$, $p$-value < 0.05). Hence, it could be inferred that there is an effective choice of amino acids for translation efficiency in the tobacco CDS. Different results were observed in tomato, where axis-1 of CA-RSCU correlated with GC3 ($r = -0.75$, $p$-value < 0.05) as well as ENc ($r = 0.37$, $p$-value < 0.05). This result reveals that directional mutation for base composition, not selection, dominates the CUB within the tomato genome. However, we did not find any remarkable trend for CA-AA and hence amino acid usage in tomato was affected by factors other than mutation and selection.

In potato, no single force could explain most of the variation in CUB, where in CA-RSCU axis-1 the highest correlation was found with CAI ($r = -0.384$, $p$-value < 0.05) and for axis-2 the highest correlation was found with GC3 ($r = -0.373$, $p$-value < 0.05). The same trend in the result was found for CA-AA. Therefore, it can be concluded that CUB in potato was influenced by different factors and mainly by mutation for base composition and translation selection.

**Conclusions**

The present study investigated the genomes of tobacco, tomato and potato, which shared the following main criteria. They uniformly use synonymous codons ending with either T or A. They show the same weak trend of CUB with a high percentage of T3/A3 and low G3/C3 content. Their nuclear genes were influenced by mutation and natural selection. Although it was not possible to obtain the required data, we are aware that the GC content in introns or intergenic sequences could be used as the reference to measure the effect of mutation bias. Finally, studying plant CUB and the factors affecting it will increase our understanding of the evolutionary history of those plants. Also, it enables determining the preferred optimal codons that help in some biotechnology applications such as gene transfer and gene editing.

**Data availability**

Data used here are available online from the websites stated in materials and methods.

**Disclosure statement**

No potential competing interest was reported by the authors.
References

[1] Mitra S, Ray SK, Banerjee R. Synonymous codons influencing gene expression in organisms. RRBC. 2016;6:57–65.

[2] Boël G, Letso R, Neely H, et al. Codon influence on protein expression in E. coli correlates with mRNA levels. Nature. 2016;529(7586):358–363. [26760206]

[3] Liu Y. A code within the genetic code: codon usage regulates co-translational protein folding. Cell Commun Signal. 2020;18(1):145.

[4] Wang L, Xing H, Yuan Y, et al. Genome-wide analysis of codon usage bias in four sequenced cotton species. PLoS One. 2018;13(3):e0194372. http://dx.doi.org/10.1371/journal.pone.0194372.

[5] LaBella AL, Opulente DA, Steenwyk JL, et al. Variation and selection on codon usage bias across an entire subphylum. PLoS Genet. 2019;15(7):e1008304.

[6] Prabha R, Singh DP, Sinha S, et al. Genome-wide comparative analysis of codon usage bias and codon context patterns among cyanobacterial genomes. Mar Genom. 2017;32:31–39. http://dx.doi.org/10.1016/j.margen.2016.10.001.

[7] Deng Y, de Lima Hedayioglu F, Kalfon J, et al. Hidden patterns of codon usage bias across kingdoms. J R Soc Interface. 2020;17(163):20190819. http://dx.doi.org/10.1098/rsif.2019.0819.

[8] Lyu X, Yang Q, Li L, et al. Adaptation of codon usage to tRNA I34 modification controls translation kinetics and proteome landscape. PLoS Genet. 2020;16(6):e1008836.

[9] Huang Y, Koonin EV, Lipman DJ, et al. Selection for minimization of translational frameshifting errors as a factor in the evolution of codon usage. Nucleic Acids Res. 2009;37(20):6799–6810.

[10] Frumkin I, Lajoie MJ, Gregg CJ, et al. Codon usage of highly expressed genes affects proteome-wide translation efficiency. Proc Natl Acad Sci U S A. 2018;115(21):E4940–E4949. https://www.pnas.org/content/115/21/E4940.

[11] Tian J, Yan Y, Yue Q, et al. Predicting synonymous codon usage and optimizing the heterologous gene for expression in E. coli. Sci Rep. 2017;7(1):1–9. http://dx.doi.org/10.1038/s41598-017-01546-0.

[12] Newman ZR, Young JM, Ingolia NT, et al. Differences in codon usage and GC content contribute to the balanced expression of TLR7 and TLR9. Proc Natl Acad Sci U S A. 2016;113(10):E1362–E1371. https://www.pnas.org/content/113/10/E1362.

[13] Romiguier J, Roux C. Analytical biases associated with GC-content in molecular evolution. Front Genet. 2017;8:16. http://dx.doi.org/10.3389/fgen.2017.00016.

[14] Wright F. The ‘effective number of codons’ used in a gene. Gene. 1990;87(1):23–29.

[15] Sharp PM, Li W-h. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Research. 1987;15(3):1281–1295.

[16] Song H, Liu J, Song Q, et al. Comprehensive analysis of codon usage bias in seven epichloé species and their peramine-coding genes. Front Microbiol. 2017;8(July):1–12.

[17] Bera BC, Virmani N, Kumar N, et al. Genetic and codon usage bias analyses of polymerase genes of equine influenza virus and its relation to evolution. BMC Genomics. 2017;18(1):652.

[18] Sablok G, Wu X, Kuo J, et al. Genomics combinational effect of mutational bias and translational selection for translation efficiency in tomato (Solanum lycopersicum) Cv. Micro-Tom. Genomics. 2013;101(5):290–295. http://dx.doi.org/10.1016/j.ygeno.2013.02.008.

[19] Novembre J. A. Letter to the editor accounting for background nucleotide composition when measuring codon usage bias. Amino Acids. 2002;19(8):1390–1394.

[20] Wickham H. 2016. Ggplot2: elegant graphics for data analysis. New York: Springer-Verlag. http://ggplot2.org.

[21] Lee BD. Python implementation of codon adaptation index. JOSS. 2018;3(30):905.

[22] Charif D, Lobry JR. 2007. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Bastolla U, Porto M, Roman HE, and Vendruscolo M, editors. Structural approaches to sequence evolution: molecules, networks, populations, biological and medical physics, biomedical engineering. New York: Springer Verlag. p. 207–232.

[23] Khandia R, Singhal S, Kumar U, et al. Analysis of nipah virus codon usage and adaptation to hosts. Front Microbiol. 2019;10(MAY):1–18.

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

[25] Zhang R, Zhang L, Wang W, et al. Differences in codon usage bias between photosynthesis-related genes and genetic system-related genes of chloroplast genomes in cultivated and wild solanum species. UMS. 2018;19(10):3142.

[26] Naya H, Zavala A, Romero H, et al. Correspondence analysis of amino acid usage within the family Bacillaceae. Biochem Biophys Res Commun. 2004;325(4):1252–1257.

[27] Kawabe A, Miyaishi NT. Patterns of codon usage bias in three dicot and four monocot plant species. Genes Genet Syst. 2003;78(5):343–352.

[28] Ossowski S, Schneeberger K, Lucas-Lledó JI, et al. The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. Science. 2010;327(5961):92–94.

[29] Camiolo S, Melito S, Porceddu A. New insights into the interplay between codon bias determinants in plants. DNA Res. 2015;22(6):461–470.

[30] Murray EE, Lotzer J, Eberle M. Codon usage in plant genes. Nucleic Acids Res. 1989;17(2):477–489.

[31] Liu H, Huang Y, Du X, et al. Patterns of synonymous codon usage bias in the model grass brachypodium distachyon. Genet Mol Res. 2012;11(4):4695–4706.