Analysis of Codon Usage Bias in the chloroplast genome of *Helianthus annuus* J-01

ShanYu Chen¹,a, Hui Zhang²,b, Xue Wang³,c, YueHua Zhang³,d, GuanHai Ruan¹,e*, Jun Ma⁴,f*

¹Institute of Crops and Nuclear Technology Utilization, Zhejiang Academy of Agricultural Sciences, Hangzhou, Zhejiang, China
²Zhejiang Agricultural Technology Extension Center, Hangzhou, Zhejiang, China
³School of Science, Jiaomusi University, Jiamusi, Heilongjiang, China
⁴Instiitude of Industrial Crops, Heilongjiang Academy of Agricultural Sciences, Harbin, Heilongjiang, China

*e-mail: csy6227862@163.com, b-e-mail: zhanghui881007@126.com, c-e-mail: wangxue_0819@163.com, d-e-mail: zhangyaohua_2008@163.com

*Corresponding Author: e-mail: ruangh@zaas.zc.cn, femail: jumhahaas@hotmail.com

Abstract. In order to determine the usage of the codon against *Helianthus annuus* chloroplast genome, this study used Codon W 1.4.2 and online software CUSP to perform neutral mapping, ENC-plot and PR2-plot analysis on the coding sequence codes of 51 genes in the *H. annuus* J-01 chloroplast genome. The results showed the GC content of *H. annuus* J-01 chloroplast genome codons was GC 1 (47.14%) > GC 2 (39.44%) > GC 3 (28.42%), indicating that the preference for synonymous codons in the chloroplast genome ends with A/U. The average ENC value is greater than 45, of which 37 are greater than 45. It shows that the *H. annuus* J-01 chloroplast genome has a weak bias. Through neutral plot analysis, ENC-plot analysis and PR2-plot analysis, it is found that the preference of *H. annuus* J-01 chloroplast genome codons is affected by selection as the main factor. The 12 optimal codons in the *H. annuus* J-01 chloroplast genome were finally determined: CUU, AUA, UAU, UAG, CAU, AAA, CCA, ACU, GCU, GAU, AGU, GGA.

1. Introduction

Chloroplasts originated from Cyanobacteria[1], which are symbiotic in primitive eukaryotes, and are semi-autonomous organelles that exist in plants. It plays a significant role in plant photosynthesis and participates in the energy supply of plants and the biosynthesis of secondary metabolites[2]. The chloroplast genome can independently perform gene transcription and translation. Due to its conservation and low evolution rate in terrestrial plants, chloroplast gene sequences have been widely used in research on plant system evolution and species identification[3]. The analysis of chloroplast genome codon bias and its influencing factors can provide references for further revealing the origin and evolution of chloroplasts.

*Helianthus annuus* belongs to the Asteraceae Helianthus is an annual herbaceous plant and is one of the four major oil crops in the world[4]. The experiment took the *H. annuus* J-01 chloroplast genome
sequence obtained by high-throughput sequencing as the research object. By analyzing the base composition, neutral plotting, ENC-plot and PR2-plot of the coding DNA sequence (CDS) of the *H. annuus* J-01 chloroplast genome, the main factors affecting the codon preference of *H. annuus* J-01 chloroplast are inferred, and the optimal codon is confirmed. It is expected to provide a scientific reference for the application and follow-up research of *H. annuus* chloroplast genome codons.

2. Materials and Methods

2.1. Sequence acquisition
Cultivated *H. annuus*, number J-01, was planted in the Science and Technology Park of Jiamusi University. Collect fresh sunflower leaves, storing the samples in dry ice, and send them to BGI for chloroplast genome sequencing. In order to reduce errors, sequences with a length of less than 300bp, containing stop codons and repetitions are excluded from the coding sequence. Finally, 51 CDS with ATG as the start codon and TAA, TGA and TAG as the stop codons were selected for subsequent analysis.

2.2. Codon related parameter calculation
Taking the selected CDS as the research object, using Codon W 1.4.2 software to analyze, to calculate the effective number of codons (ENC) and relative synonymous codon usage (RSCU); calculate the content of nucleotides T, C, A and G at the third position of the codon and mark them as T3, C3, A3 and G3, respectively. Using the CUSP online software analysis of EMBOSS explorer to calculate the GC content at the 1, 2 and 3 positions of the codon and the overall GC content, denoted as GC1, GC2, GC3 and GC_all, respectively.

2.3. Neutral drawing analysis
Calculating the average value according to the obtained GC1 and GC2 content, record it as GC12, draw a scatter plot with GC12 as the ordinate and GC3 as the abscissa. If GC12 and GC3 are significantly correlated, and the regression coefficient is close to 1, the base composition at the three positions of the codon is similar, and the codon preference is mainly affected by the mutation. On the contrary, it shows that the base composition of the 1, 2 and 3 codons is quite different, the GC content of the genome is highly conservative, and its preference is greatly affected by choice.

2.4. ENC-plot drawing analysis
ENC-plot drawing analysis is the effective codon number drawing analysis, reflecting the relationship between base composition and codon bias. Drawing a scatter chart with GC3 as the abscissa and ENC as the ordinate, and add a standard curve in the chart: ENC=2+GC3+29/[GC32+(1-GC3)2]. The standard curve shows the linear relationship between GC3 and ENC under no selection pressure. Closing to the standard curve means that the codon usage bias is only affected by mutations; on the contrary, in addition to mutations, there are other factors, especially selection.

2.5. PR2-plot analysis
PR2-plot analysis is also called parity preference analysis, which explores whether the mutations of base A, T, C, and G in the third base of each gene codon are balanced. Using A3/(A3+T3) as the ordinate and G3/(G3+C3) as the abscissa to draw a scatter graph. The centre of the graph follows the PR2 principle. The vector emitted from the centre point (A=T, C=G) indicates the degree and direction of the base shift.

2.6. Optimal codon screening
Sort the ENC values of 51 CDSs from low to high, and select 10% genes at each end (5 CDSs before and after each), and define them as high expression group and low expression group. Calculate the RSCU difference between the high expression group and the low expression group based on the obtained RSCU value, and record it as ΔRSCU. The codons with ΔRSCU≥0.08 between the two groups are
regarded as high-expression superior codons, and the codons with RSCU>1 are defined as high-frequency codons, and when two conditions are met at the same time, it is determined as the optimal codon.

3. Results and discussion

3.1. Genome codon composition

The 51 chloroplast genome CDS sequences selected were analyzed (Table 1). The results showed that the average content of the 1, 2 and 3 codons was different, from largest to smallest, they were GC1, GC2, and GC3, and the GC content was 47.14%, 39.44% and 28.42%, respectively. The average content of GCall is 38.33%, and the content of GC3 is the lowest, indicating that the last base of the codon in the *H. annuus* J-01 chloroplast genome prefers to use A/T(U). The value range of ENC is 39.13~59.17, and the average value is 47.63. Using ENC>45 as the criterion for weak codon usage preference, there are 37 genes with ENC>45, which shows that the *H. annuus* J-01 chloroplast genome has a weak codon usage preference.

### Table 1. GC content and ENC value of different positions of codon

| gene    | GC1  | GC2  | GC3  | GCall | ENC  | gene    | GC1  | GC2  | GC3  | GCall | ENC  |
|---------|------|------|------|-------|------|---------|------|------|------|-------|------|
| rps12   | 50.78| 47.66| 30.47| 42.97 | 47.97| petA    | 52.65| 37.38| 30.84| 40.29 | 50.97|
| psbA    | 50.00| 43.50| 33.05| 42.18 | 41.16| rps18   | 34.31| 42.16| 25.49| 33.99 | 44.76|
| matK    | 38.52| 34.33| 27.94| 33.60 | 46.82| rpl20   | 37.80| 45.67| 27.56| 37.01 | 46.96|
| rpoB    | 50.52| 37.32| 27.80| 38.55 | 48.50| clpP    | 59.39| 38.07| 29.95| 42.47 | 49.38|
| rpoC1   | 50.29| 36.52| 26.23| 37.68 | 48.88| psbB    | 53.83| 45.97| 30.45| 43.42 | 47.88|
| rpoC2   | 46.52| 37.42| 30.45| 38.13 | 50.93| petB    | 49.07| 41.67| 27.31| 39.35 | 43.52|
| rps2    | 43.46| 43.04| 27.00| 37.83 | 45.89| petD    | 51.55| 39.13| 27.33| 39.34 | 39.25|
| atpI    | 50.40| 37.90| 28.23| 38.84 | 46.54| rpoA    | 44.35| 32.14| 29.76| 35.42 | 51.62|
| atpF    | 46.49| 32.43| 32.97| 37.30 | 51.82| rps11   | 54.74| 55.47| 27.01| 45.74 | 50.38|
| atpA    | 56.19| 59.49| 25.34| 40.34 | 46.88| rps8    | 41.48| 40.74| 28.15| 36.79 | 48.37|
| psbD    | 52.26| 43.22| 32.42| 46.66 | 44.10| rpl14   | 56.10| 37.40| 26.83| 40.11 | 45.15|
| psbC    | 53.38| 46.20| 29.54| 43.04 | 43.82| rps3    | 46.12| 33.33| 23.74| 34.40 | 43.16|
| rps14   | 44.55| 47.52| 33.66| 41.91 | 39.24| rpl22   | 39.35| 40.00| 23.87| 34.41 | 39.13|
| psaB    | 48.03| 43.40| 29.25| 40.23 | 46.17| ycfI    | 34.94| 27.43| 26.73| 29.70 | 46.26|
| psaA    | 52.80| 43.15| 33.80| 43.25 | 51.02| ycfF    | 37.34| 38.46| 32.54| 39.45 | 59.17|
| ycf3    | 47.34| 38.46| 32.54| 39.45 | 59.17| ndhA    | 45.33| 39.84| 18.96| 34.71 | 41.24|
| rps2    | 53.57| 39.60| 28.71| 40.59 | 52.61| ndhB    | 41.92| 35.93| 25.75| 34.53 | 48.32|
| ndhJ    | 50.94| 37.11| 30.82| 39.62 | 43.40| ndhG    | 45.76| 33.90| 28.81| 36.16 | 53.13|
| ndhK    | 43.81| 44.25| 25.37| 38.20 | 51.26| ndhE    | 42.16| 34.31| 18.63| 31.70 | 43.98|
| ndhC    | 45.45| 33.06| 24.79| 34.44 | 48.97| ccsA    | 32.82| 37.15| 24.77| 31.58 | 49.45|
| atpE    | 52.99| 41.04| 25.37| 39.80 | 51.30| ndfF    | 36.96| 35.75| 24.60| 32.44 | 45.22|
| atpB    | 55.71| 41.68| 29.06| 42.15 | 48.21| rps7    | 53.21| 46.15| 23.72| 41.03 | 44.31|
| rbcL    | 57.00| 43.00| 30.86| 43.62 | 47.42| ndhB    | 41.29| 38.55| 32.09| 37.31 | 48.60|
| accD    | 42.20| 35.76| 30.35| 36.11 | 50.67| ycf2    | 41.89| 34.24| 36.91| 37.68 | 52.81|
| ycf1    | 44.32| 40.54| 30.27| 38.38 | 51.04| rpl2    | 50.18| 47.27| 33.45| 43.64 | 56.40|
| cemA    | 39.13| 28.26| 30.87| 32.75 | 49.80| Average | 47.14| 39.44| 28.42| 38.33 | 47.63|

Note: GCall represents the average number of codon positions.

The correlation between the codon parameters of the *H. annuus* J-01 chloroplast genome was analyzed (Table 2). It showed that GC1 and GC2, GC1 and GCall, GC2 and GCall, GC3 and GCall were extremely significantly correlated, and GC3 was not significantly associated with GC1 and GC2, indicating the base composition of the 1 codon and the 2 codon are similar, but is different from the base composition of the 3 codon is different. GC3 is extremely significantly related to ENC, while GC1 and GC2 are not significantly related to ENC, indicating that the GC content of the 3 codon has an impact on codon preference.
Table 2 Correlation analysis of GC content and ENC value of different position codon numbers

|       | GC1 | GC2 | GC3  | GCall | ENC   |
|-------|-----|-----|------|-------|-------|
| GC2   | 0.395** |     |      |       |       |
| GC3   | 0.212 | 0.122 |      |       |       |
| GCall | 0.835** | 0.745** | 0.502** |       |       |
| ENC   | 0.074 | -0.12 | 0.407** | 0.116 |       |
| N     | -0.153 | -0.298* | 0.269 | -0.144 | 0.162 |

Note: * indicates that the correlation has reached a significant level (P < 0.05); ** indicates that the correlation has reached a extremely significant level (P < 0.01)

3.2. Neutral drawing analysis

The result of neutral mapping analysis of codon usage in H. annuus J-01 chloroplast genome (Figure 1) showed that the value range of GC3 was between 0.186 and 0.369, and the value range of GC12 was between 0.312 and 0.551. The correlation analysis of GC12 and GC3 shows that the correlation coefficient is 0.202, the correlation is not significant, and the regression coefficient is 0.1468. It shows that the 1 and 2 base of the H. annuus J-01 chloroplast genome have little correlation with the 3 base, and the mutation has a weak influence on the formation of codon bias. The choice may have a more significant impact on the formation of codon preference.

3.3. ENC-plot analysis

The results of the association analysis between GC1 and ENC (Figure 2) showed that some genes are located near the standard curve, and the actual value of ENC of these genes is close to the expected value of ENC. Most of the genes are located below the standard curve, indicating that among the influencing factors of their codon preference, the choice is dominant.

3.4. PR2-plot analysis

PR2-plot was used to analyze the relationship between the bases at position 3 of the codons of some amino acids of each encoding gene in the H. annuus J-01 chloroplast genome (Figure 3). The results show that the genes are not evenly distributed in the region, and most of them fall on the lower right of the figure, indicating that in terms of base usage frequency, T>A and G>C. There is an imbalance in the use of A and T, C and G, and the four bases are not evenly distributed on the plane of PR2, which indicates that the codon usage pattern of H. annuus J-01 chloroplast genome is affected by a combination of many factors, not just mutations.
3.5. Optimal codon analysis

The results (Table 3) showed that there were 34 high-frequency codons with RSCU>1 and 31 high expression codons with ΔRSCU≥0.08, of which 7 ends with U, 5 ends with A, 9 ends with C, and 10 ends with G. Combine high-frequency codons with high expression codons to obtain a total of 12 optimal codons. They are CUU, AUA, UAG, CAU, AAA, CCA, ACU, GCU, GAU, AGU, GGA, of which 7 ends with U, 4 ends with A, and 1 end with G.

| Codon | High expression gene RSCU number | Low expression gene RSCU number | ΔRSCU | Codon | High expression gene RSCU number | Low expression gene RSCU number | ΔRSCU |
|-------|----------------------------------|----------------------------------|-------|-------|----------------------------------|----------------------------------|-------|
| Phe   | UUU                             | 91                               | 1.33  | Tyr   | UAU*                            | 91                               | 1.448 | 0.11  |
|       | UUC*                            | 86                               | 0.67  | UAC   | 24                               | 0.552 | 0.11  |
| Leu   | UUA                             | 67                               | 1.742 | 50     | 2.31                            | -0.568 | 2       | 1.2   |
|       | UAG                             | 77                               | 1.236 | 25     | 1.392                           | -0.156 | 2       | 1.2   |
|       | CUG*                            | 82                               | 1.312 | 28     | 1.084                           | 0.228 | 61      | 1.56  |
|       | CUC**                           | 28                               | 0.482 | 5      | 0.154                           | 0.33  | 18      | 0.44  |
|       | CUA                             | 48                               | 0.836 | 28     | 0.918                           | -0.082 | 75      | 1.28  |
|       | CUG*                            | 26                               | 0.388 | 4      | 0.144                           | 0.244 | 38      | 0.72  |
| Ile   | AUU                             | 99                               | 1.29  | 75     | 1.77                            | -0.48  | 130     | 1.402 |
|       | AUA                             | 58                               | 0.592 | 18     | 0.344                           | 0.248 | 45      | 0.598 |
|       | AUG                             | 58                               | 1      | 30     | 1                               | 0.08  | 62      | 0.378 |
|       | AUU*                            | 59                               | 1.116 | 25     | 0.804                           | 0.242 | 120     | 1.622 |
| Met   | AUG                             | 58                               | 1      | 30     | 1                               | 0      | 62      | 0.378 |
|       | AUG*                            | 58                               | 1      | 30     | 1                               | 0      | 62      | 0.378 |
| Val   | GUU                             | 36                               | 0.98  | 27     | 1.846                           | -0.866 | 123     | 1.61  |
|       | GUC***                          | 23                               | 0.642 | 4      | 0.144                           | 0.502 | 30      | 0.39  |
|       | GUA                             | 30                               | 1.382 | 33     | 1.762                           | -0.38  | 100     | 1.484 |
|       | GUG***                          | 27                               | 0.996 | 4      | 0.254                           | 0.742 | 54      | 0.516 |
| Ser   | UCU                             | 67                               | 1.318 | 27     | 1.242                           | 0.076 | 21      | 1.322 |
|       | UCC                             | 47                               | 1.052 | 16     | 1.308                           | -0.256 | 11      | 0.678 |
|       | UCA                             | 57                               | 1.414 | 19     | 1.652                           | -0.238 | 5       | 0.6   |
|       | UCG*                            | 30                               | 0.708 | 3      | 0.272                           | 0.436 | 47      | 1      |
|       | UCA                             | 48                               | 1.594 | 22     | 1.402                           | 0.192 | 45      | 1.34  |
| Thr   | ACC                             | 22                               | 0.768 | 11     | 0.644                           | 0.124 | 14      | 0.166 |
|       | ACC*                            | 22                               | 0.768 | 11     | 0.644                           | 0.124 | 14      | 0.166 |
|       | ACA                             | 40                               | 1.254 | 23     | 1.604                           | -0.35  | 61      | 1.25  |
|       | ACG                             | 19                               | 0.384 | 2      | 0.346                           | 0.038 | 34      | 0.59  |
| Ala   | GCU***                          | 46                               | 1.608 | 40     | 1.466                           | 0.142 | 45      | 1.054 |
|       | GCC**                           | 22                               | 0.9    | 7      | 0.28                            | 0.62  | 9       | 0.192 |
|       | GCA                             | 31                               | 0.898 | 28     | 1.662                           | -0.764 | 69      | 1.784 |
|       | GCC                             | 13                               | 0.592 | 9      | 0.59                            | 0.002 | 40      | 0.974 |

Note: * means ΔRSCU ≥ 0.08, ** means ΔRSCU ≥ 0.3, *** means ΔRSCU ≥ 0.5. The underlined codons indicate that its genome RSCU>1, and the bolded codons indicate that it is the optimal codon.
4. Conclusions

*H. annuus* J-01 chloroplast genome CDS prefers to use A/T bases, and the GC content of the bases at each position of the codon is different, and the order is GC1>GC2>GC3. The number of effective codons ranges from 20 to 61. The larger the value, the weaker the preference. There are 37 CDSs with ENC>45, indicating that the *H. annuus* J-01 chloroplast genome has a weaker codon preference. According to research, the reasons affecting the formation of codon preference are complex, including factors such as mutation pressure, natural selection[5], gene length, tRNA abundance[6] and GC content[7]. However, selection and mutation are the main reasons that force the formation of codon preference. In this study, neutral analysis, ENC-plot, and PR2-plot analysis were chosen to explore the reasons for the formation of *H. annuus* J-01 chloroplast genome codon preference. The results show *H. annuus* J-01 chloroplast genome codons are mainly affected by selection pressure, and mutation pressure is less affected. The combination of high frequency and high expression codons screened out 12 optimal codons, of which 7 ended with U and 4 ended with A. The characteristics of the optimal codon composition were similar to those of most species. This research systematically explored the factors affecting the codon preference of *H. annuus* J-01 plants and conducted a comprehensive analysis of the codon preference of *H. annuus* J-01 chloroplast genome to provide a scientific basis for subsequent research on heterologous expression of *Helianthus* exogenous gene codon modification and chloroplast genetic engineering.

Acknowledgement

Zhejiang Province Key R&D Program (2019C02004); The Six-Party Scientific and Technological Collaboration Plan (2019SNLF003); "Fourteenth Five-Year Plan" Major Special Project for Breeding New Agricultural Varieties in Zhejiang Province (Project for Breeding New Oilseed Varieties).

References

[1] Daniell H, Lin CS, Yu M, Chang WJ. (2016) Chloroplast genomes: diversity, evolution, and applications in genetic engineering. J. Sci. Genome biology, 17: 134-141.

[2] Zhang R, Zhang L, Wang W, et al. (2018) Differences in Codon Usage Bias between Photosynthesis-Related Genes and Genetic System-Related Genes of Chloroplast Genomes in Cultivated and Wild Solanum Species. J. Sci. International journal of molecular sciences, 19: 31-42.

[3] Zoschke R, Bock R. (2018) Chloroplast Translation: Structural and Functional Organization, Operational Control, and Regulation. J. Sci. The Plant cell, 30: 745-770.

[4] Vangelisti A, Mascagni F, Giordani T, et al. (2019) Arbuscular mycorrhizal fungi induce the expression of specific retrotransposons in roots of sunflower (*Helianthus annuus* L.). J. Sci. PloS one, 14: 78-86.

[5] Wang Z, Xu B, Li B, et al. (2020) Comparative analysis of codon usage patterns in chloroplast genomes of six Euphorbiaceae species. J. Sci. PeerJ, 8: 114-121.

[6] Szövényi P, Ullrich KK, Rensing SA, et al. (2017) Selfing in Haploid Plants and Efficacy of Selection: Codon Usage Bias in the Model Moss *Physcomitrella patens*. J. Sci. Genome biology and evolution, 9: 1528-1546.

[7] LaBella AL, Opulente DA, Steenwyk JL, et al. (2019) Variation and selection on codon usage bias across an entire subphyllum. J. Sci. PLoS genetics, 15: 234-246.