Codon usage bias analysis of 22 *Anemone*

Zeyue Yu*

College of Life Sciences, Qingdao University, Qingdao, Shandong, 266071, China

* Corresponding Author Email: yuzeyue190013421@126.com

Abstract. Codon preference analysis for the *rbcL* gene of *Ranunculaceae* plants remains to be completed, and for this purpose we selected 10801 codons from 23 *Ranunculaceae Juss.* plants for analysis. Among them, 30 codons had relative synonymous codon usage values greater than or equal to 1. The codon endings were preferred to A and U, among which U was more preferred. Neutrality plot analysis, ENC-GC3s plot analysis and PR2 bias plot analysis showed that codon preference in *Anemone L.* was mainly influenced by selection pressure. Cluster analysis showed that a phylogenetic tree could be constructed for *Anemone L.* based on relative synonymous codon usage values. Comparison with codon usage frequencies in model organisms indicated that *Arabidopsis thaliana* and *Saccharomyces cerevisiae* are more suitable vectors for exogenous expression of the *rbcL* gene in *Anemone L.* plants. These results help to elucidate the molecular biological characteristics of *Ranunculaceae Juss.* plants for further exploitation of *Ranunculaceae Juss.* plant resources in areas such as synthetic biology.

Keywords: Anemone*, rbcL*, Codon usage bias, cluster analysis, RSCU.

1. Introduction

The genetic central dogma of molecular biology can be expressed as the transcription of DNA into mRNA, which is then translated into proteins. During translation, each tRNA provides a direct, specific link between a nucleotide triplet (i.e. a codon) and the corresponding amino acid. The ribosome is the translational molecular machine that houses the tRNAs and mRNAs. Studies of the genetic codons have shown that 61 codons, which encode the standard 20 amino acids. At the same time, there are 3 codons which are translation termination signals. The genetic code is almost universal. We can get from it that almost all organisms use the same codons for specific amino acids. As 18 of the 20 amino acids are encoded by multiple synonymous codons, the genetic code is called 'parsimony'. Because synonymous mutations do not affect the identity of the amino acid encoded, they were originally thought to have no effect on protein function or the protein itself. However, as organisms evolve and adapt, the appearance of synonymous codons is not "neutral" (i.e. randomly selected). For the genome of a given species, some synonymous codons occur frequently, others rarely or not at all. Different species appear to be specific in their codon usage. This phenomenon is known as codon preference [1].

There are several possible sources of codon preference, including changes in tRNA availability, changes in global and local G + C content, mutational bias, changes in mRNA stability and transcription rate, and changes in codon, anticodon interactions. Often used in research is selection pressure mutation pressure versus selection pressure [2][3][4][5].

With the maturity of DNA sequencing and the development of technologies such as bioinformatics [6], relevant indices for studying codon preferences have gradually been proposed, such as the Codon Adaptation Index (CAI), the Effective Codon Count (ENC) [3][4]. Comparisons of codon preferences based on model organisms are also commonly used to study codon preferences The study of codon preferences can help us understand certain genes in certain species and apply them to biomanufacturing. The study of codon preferences can help us better understand certain genes in certain species and apply them to biomanufacturing [5][7][8].

Ranunculaceae Juss, belongs to *Plantae*, *Angiosperms*, *Dicotyledons*, *Buttercups*, among which *Anemone* has rich medicinal resource value [9]. At present, there is still a gap in the work on codon preference analysis for *Ranunculaceae Juss.* plants such as *Anemone*, and the related field of analysis has to be filled. I hypothesize that *Anemone* has their own unique codon preferences and specific data

can be derived using software and analyses including CodonW. This preference is reflected in the reasons for its formation, base selection tendencies, etc. In order to study codon preference in *Ranunculaceae Juss.* we selected the *rbcL* gene sequences of 23 *Ranunculaceae Juss* and *Anemoclema glaucifolium* for codon preference analysis, explored the reasons for the formation of codon preference in the study subjects mainly *Anemone*, and on the basis of codon preference analyses, a phylogenetic study of *Ranunculaceae Juss.* was carried out.

2. Materials and methods

2.1. Acquisition of rbcL sequences from Anemone and other Ranunculaceae plants

The *rbcL* gene sequences in the corresponding chloroplast genomes were obtained from the nucleic acid sequence libraries of the NCBI public database (https://www.ncbi.nlm.nih.gov/) and collated into the prescribed fasta format files for subsequent codon preference analyses [6].

2.2. Codon composition analysis

Codon usage bias as well as structural composition of 23 *rbcL* sequences were analyzed using CodonW software (http://codonw.Sourceforge.Net).

Effective number of codon (ENC). The ENC value ranges from 20 to 61, where 20 indicates that only one codon is used per amino acid and 61 indicates that each codon is used equally. The lower the value, the stronger the codon preference and vice versa, ENC can reflect the degree of preference for unbalanced use of synonymous codons in a codon family and is the most informative parameter for assessing the overall codon preference of a gene. Typically, highly expressed genes have a large codon usage bias and thus their ENC values are smaller; low expressed genes contain more types of rare codons and their ENC values are smaller. Therefore, the relative level of expression of endogenous genes can be determined by comparing ENC values [10][11].

Codon adaptation index (CAI)**.** For a given gene, the CAI is the fitness coefficient of all codons coding for the protein relative to the case where the optimal codon is used for this gene. To calculate this value, a table of optimal codons for highly expressed genes in the corresponding species should be provided as a reference. The CAI value is between 0 and 1, the higher the value, the more adaptive the gene is and the higher the expression level is [3][12]. Codon bias index (CBI). This indicator responds to the component of highly expressed superior codons in a gene. Relative synonymous codon usage (RSCU). RSCU is defined as the number of times a particular synonymous codon is used as the numerator and the number of times that codon is expected to occur as the denominator. Where the number of predicted occurrences is the average number of times all codons for the amino acid encoded by that codon are used. The results were plotted and analyzed using R [4][12]. The values of each sequence CG, CG_1, CG_2 and CG_3 were obtained by CUSP calculation in EMBOSS explorer (https://www.Bioinformatics.nl/emboss explorer/). Kendall based correlation analysis between the parameters CG, CG_1, CG_2, CG_3 and ENC was analyzed using the software Origin 2021.

2.3. Neutrality and ENC-GC3s plot analysis

Neutral plotting analysis allows for the analysis of factors influencing codon usage preference. Taking GC_3 as the horizontal coordinate and CG_{12}, the average value of GC_1 and GC_2, as the vertical coordinate, a two-dimensional scatter plot was drawn using Origin 2021 and linearly fitted to both. When the slope of the fitted curve is close to 1, there is little difference in the base composition of each codon site, and the preference is mainly affected by mutation; when the slope is close to 0, there is a large difference in the base variation pattern between codon site 3 and sites 1 and 2, and the preference is mainly affected by selection pressure [4].

The ENC-GC3 scatter plot was plotted by Origin 2021 using GC3 as the horizontal coordinate and ENC as the vertical coordinate. The ENC expectation value was used as the standard curve, and the scatter points were fitted accordingly. The function for calculating the ENC expectation value is demonstrated below:

$$y = 2 + x + \frac{29}{x^2+(1-x)^2} \quad \textbf{(1)}$$

, where y represents the ENC expectation and x represents CG3.

If the distribution of correspondence points is around the expected curve, the codon preference is only constrained by mutational pressure. If the distribution of counterpoints is below or far from the expected curve, some other factor, such as natural selection, should play a key role in the formation of codon preferences [8].

2.4. PR2 bias plot analysis

Using the parity properties of nucleic acid sequences, the extent to which selection and mutation affect codon preference can be analyzed based on the amount of A_3, T_3, C_3 and G_3 at each codon analyzed [14].The PR2-plot scatter plot was plotted using origin 2021 with $A_3/(A_3 + T_3)$values in horizontal coordinates and $G_3/(G_3 + C_3)$ values in vertical coordinates. If all the scatter points are distributed around the intersection point (0.5,0.5), only the mutation pressure has an effect on codon preference; if the scatter points are distributed away from the intersection point (0.5,0.5), the selection pressure has a stronger effect on codon preference than the mutation pressure [12].

2.5. Phylogenetic and Cluster analysis

The 23 *rbcL* gene sequences were processed and analyzed using MEGA. X software, and a phylogenetic tree was generated using the neighbour-joining (NJ) method . The relevant parameters are shown below:

Test of phylogeny: bootstrap method, No. of bootstrap replications: 500;

Substitution model: p-distance.

Gaps/Missing data treatment: partial deletion, site coverage cutoff: 50%

Cluster analyses were also carried out on 23 plant species based on the RSCU values calculated by CodonW [5][10]. I used averaging method for the clustering operation in this cluster analysis.

2.6. Comparative analysis of codon usage frequency

The relative frequency ratio of codon usage is an important parameter to assess the degree of difference in codon usage preference among different organisms. When the ratio is 0.5-2.0, species codon preference differences are considered to be small. The genomic codon usage frequencies of *Escherichia coli*, *Saccharomyces cerevisiae, Arabidopsis thaliana*, *Zea mays*, and *Nicotiana tabacum* were found from the codon usage database (http://www.kazusa.or.jp/codon/). The average codon usage bias of the genes associated with 23 species of *Anemone* and *Anemoclema* plants were determined by EMBOSS explorer (https://www.Bioinformatics.nl/emboss explorer/) in CUSP calculation was obtained. Frequency ratios were calculated and plotted using Origin 2021 [11].

3. Results and discussion

3.1. RbcL gene information

Information on the *rbcL* gene was obtained from the public database NCBI for 23 species *Anemone* and *Anemoclema* plants (**Table 1)**.

Table 1. Information on *rbcL* genes of 23 species of *Anemone* and *Anemoclema* plants

Species	Login Number	CDS location
Anemone reflexa	NC 054280.1	56412-57839
Anemone taipaiensis	NC 050873.1	60854-62281
Anemone raddeana	NC 041526.1	55841-57268
Anemone trullifolia	NC 039456.1	54743-56170
Anemone flaccida	MT317179.1	54690-56117
Anemone altaica	NC 064389.1	56481-57908
Anemone tomentosa	NC 039451.1	56901-58328
Anemone nemorosa	NC 066446.1	56442-57869
Anemone narcissiflora	NC 045879.1	55698-57125
Anemone alpina	NC 063528.1	57455-58882
Anemone rivularis var.flore-minore	OR251125.1	57698-59125
Anemone vitifolia	NC 056381.1	57644-59071
Anemone coronaria	NC 082177.1	56170-57597
Anemone hortensis	NC 066445.1	55714-57141
Anemone occidentalis	NC 063529.1	57725-58591
Anemone shikokiana	NC 069812.1	55627-57054
Anemone turczaninovii	MK860686.1	57656-59083
Anemone henryi	NC 039465.1	55182-56609
Anemone patens subsp.mul	MN025345.1	57392-58819
Anemone hepatica var. japonica	NC 045878.1	56779-58206
Anemone hepatica var. asiatica	NC 045910.1	56311-57738
Anemone pulsatilla subsp.	MN025344.1	57132-58559
Anemoclema glaucifolium	MH205609.1	56293-57720
Thalictrum baicalense	MW133265.1	56461-57888

3.2. Parametric analysis

The codon-related bias indexes of the selected *rbcL* gene sequences are shown in Figure 1. The average GC content of the 23 *rbcL* sequences was 0.444, indicating that the gene prefers A and T bases. As far as the third position of the codon is concerned, the content of base T is the largest, with an average content of 0.477. The content of base A is the second largest, with an average content of 0.388, indicating that the codon of the gene prefers to end with A and T. The codon is not a base of the gene, but a base of the gene, with an average content of 0.388.

We analyzed the CAI values of the 23 sequences. The range of CAI values was 0.236-0.261, with an average value of 0.254, indicating that the gene is low adaptive and has low expression levels in the 23 selected species. The calculated CBI values for the 23 species ranged from 0.023-0.076 with a mean value of 0.062.

The maximum GC content was found in *Anemone occidentalis* (0.453) and the minimum was found in Anemone *flaccida* (0.435). The average content of GC_1, GC_2, and GC_3 was 0.574, 0.429, and 0.328, respectively, with GC_1 being the largest, and GC_3 the smallest. The mean of the ENC values was 50.584. where *Anemone occidentalis* (54.26) had the largest ENC value and *Anemone flaccida* (49.3) had the smallest.

From the correlation analysis in Figure 2, it can be seen that there was a significant correlation between CG_2 and CAI index ($p \leq 0.001$), which indicated that the content of CG_2 had a significant effect on CAI, and that there was a significant correlation between CG content and CG_3 ($p \leq 0.001$), which indicated that CG_3 could influence the CG content to a certain extent. In addition, there was a significant correlation between CG_2 and CBI ($p \leq 0.01$), which can be considered that the value of CBI can be influenced by CG_2.

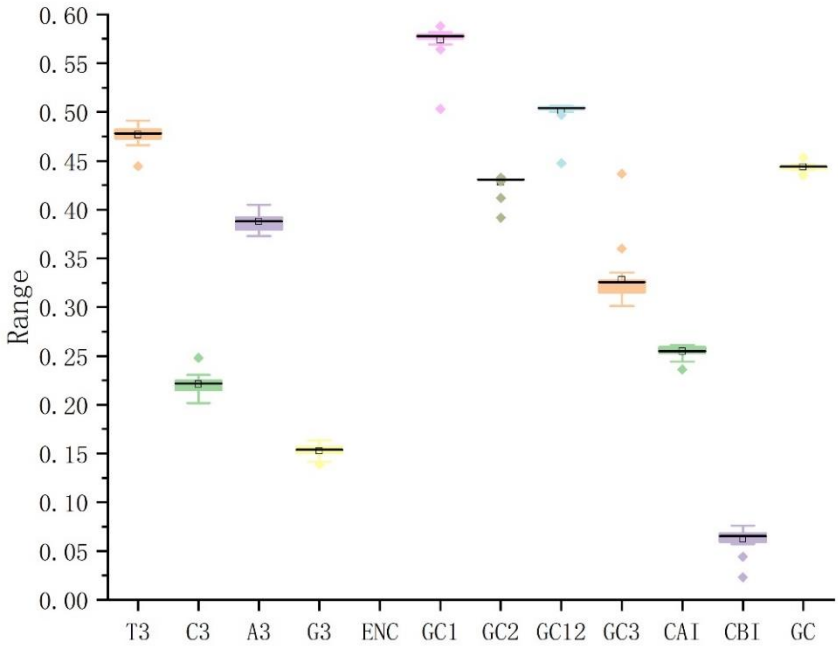


Figure 1. Box plot of codon usage indicators of 23 species. The top of the box plot must include 25 per cent of the high value data, the bottom must include 25 per cent of the low value data, and the middle line of the box is the median of all data.

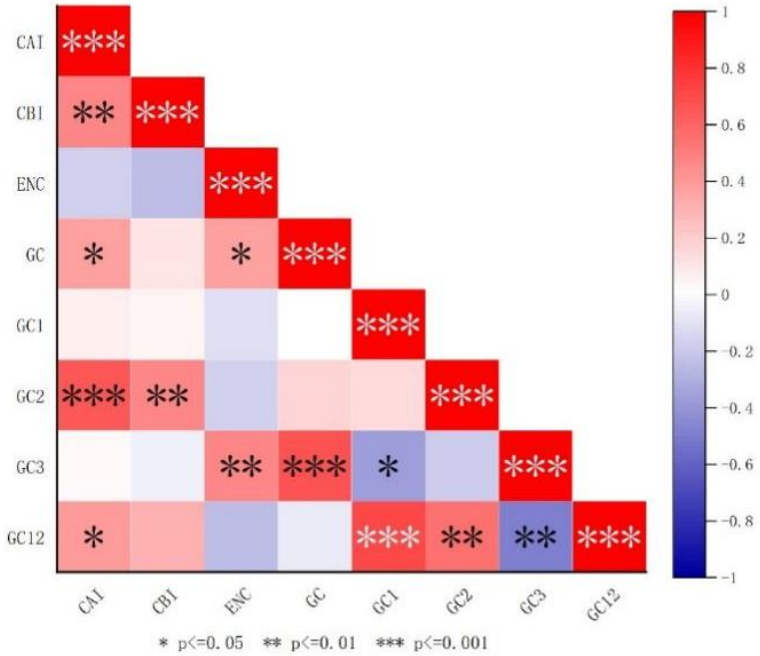


Figure 2. Heatmap of Kendall correlation analysis for different codon usage metrics. Different colours indicate correlation between variables

3.3. Codon RSCU analysis

The *rbcL* genes of 23 species of *Anemone* and *Anemoclema* plants include a total of 10,801 codons. As shown in Figure 3, methionine and tryptophan correspond to only one amino acid coded. Leucine is the most abundant, encoded by six codons. The remaining amino acids are encoded by 2-4 codons. We can see that there are 30 codons' RSCU values greater than or equal to 1, i.e., they are used frequently. Out of these 30 codons, a total of 21 end with A and U. The most frequently used codon CGU ends with U, indicating that the codon ends with a preference for A and U, with the U preference being greater.

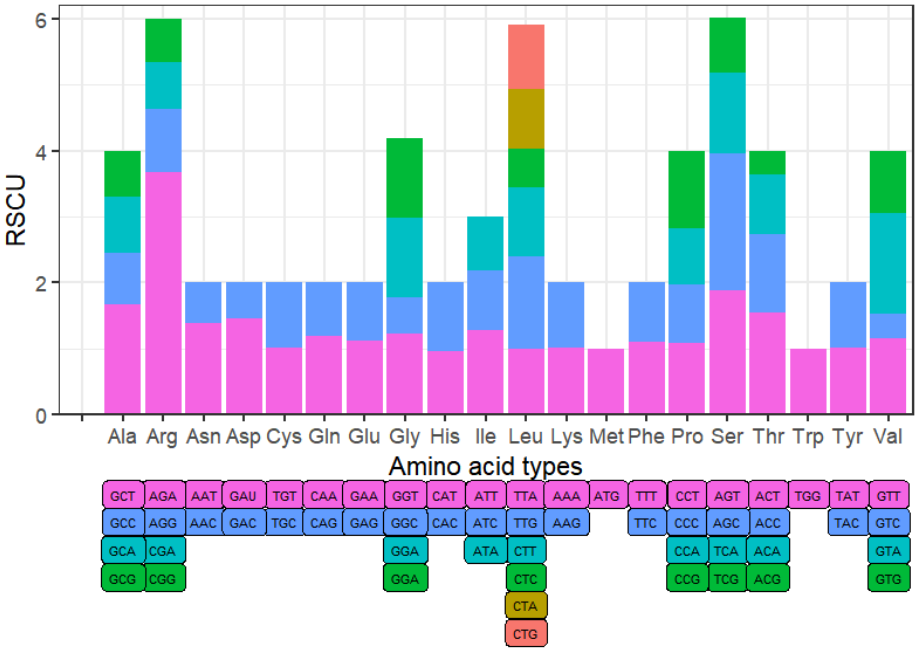


Figure 3. RSCU analysis of *Anemone rbcL* gene. Different colour are used to indicate different RSCU values

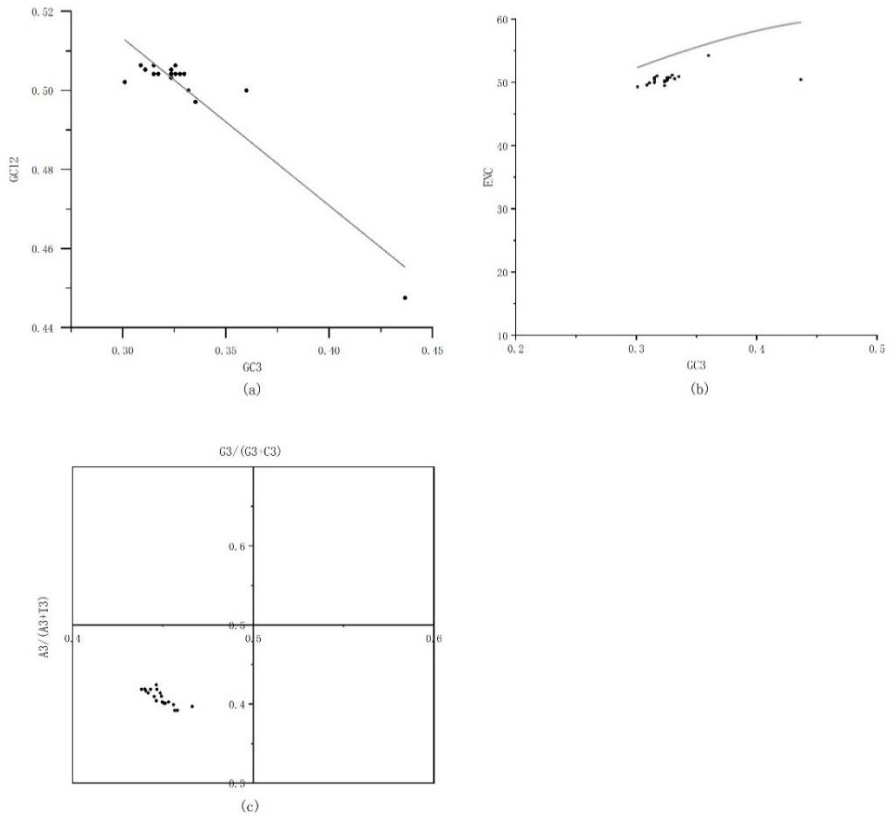


Figure 4. (a) Neutrality plot analysis (b) ENC-GC3s plot analysis (c) PR2 bias plot analysis

3.4. Neutrality plot analysis

After linear fitting, the absolute value of the slope of the curve tends to be close to 0, indicating that the base variation patterns of codon site 3 and sites 1 and 2 are more different, meanwhile, in addition, the R^2 value of the standard curve ranges from 0.866. There is no significant correlation between the GC_{12} value and the GC_3 value ($p>0.05$). It indicates that codon preference is mainly affected by selection pressure (Figure 4a).

3.5. ENC-GC3s plot analysis

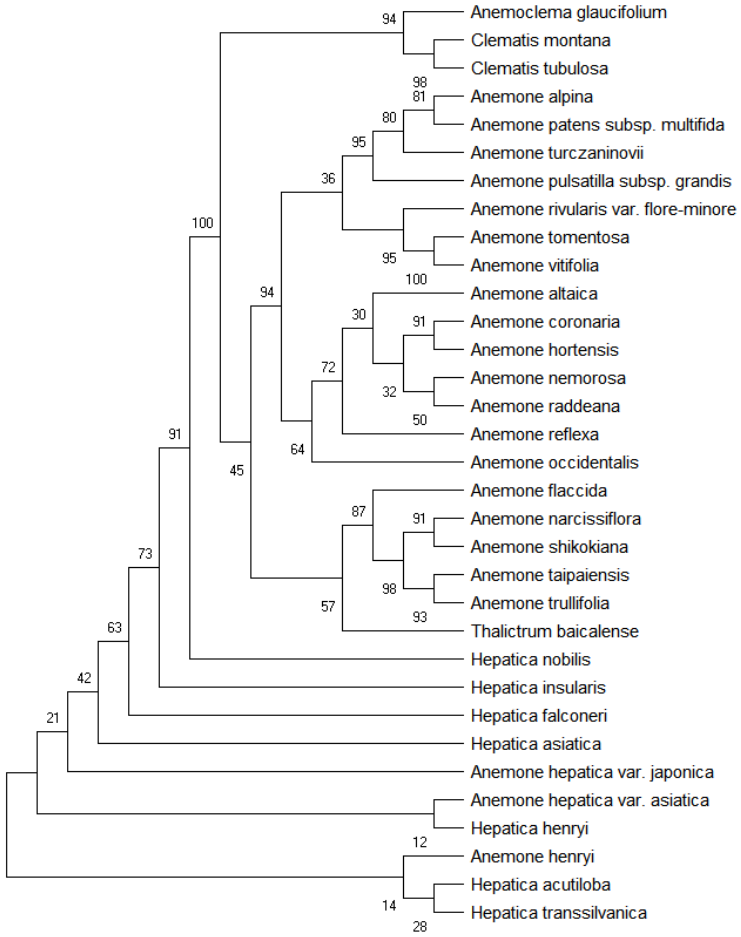
The ENC-plot results showed that all the loci were distributed below the standard curve, indicating that natural selection has a major influential role on codon preference (Figure 4b).

3.6. PR2 bias plot analysis

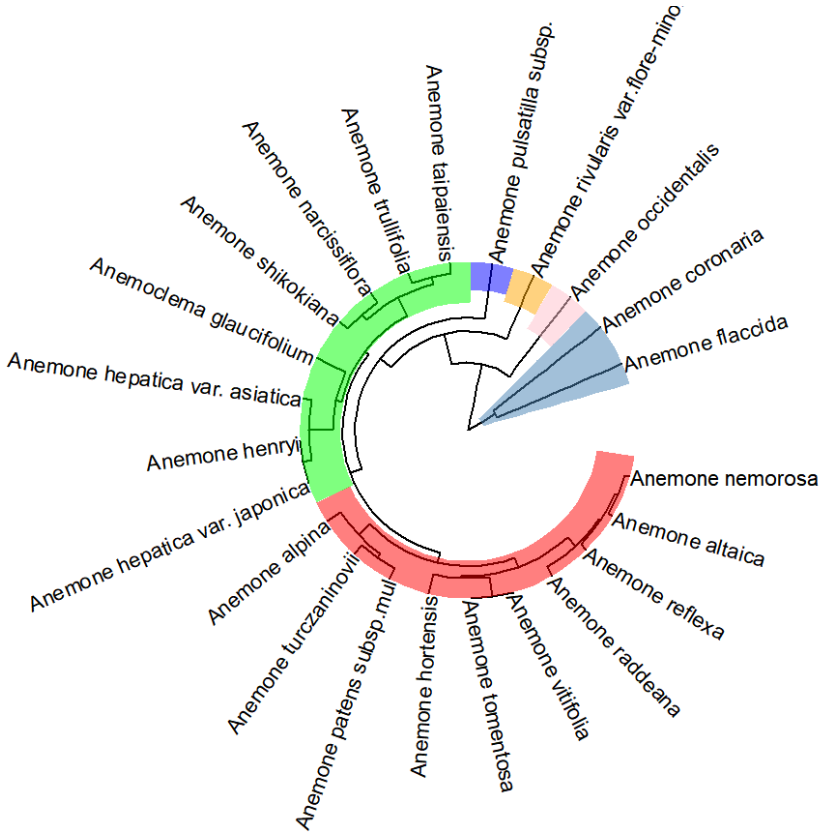
The results showed that all 23 *rbcL* genes were located in the third quadrant, indicating that the codons of these 23 genes have significant preference for the third base, as shown by the use of the third base, T is larger than A and C is larger than G. The results also showed that selection pressure is important for the codon preference of *rbcL* genes of *Anemone.* (Figure 4C)

3.7. Phylogenetic and cluster analyses

The selected plants were analyzed by clustering using CDS sequences and RSCU values of all codons, respectively, and the results are shown in Figures 5 and Figures 6. Although there were differences in the internal branching between the NJ tree and the clustering diagram, there were also consistent clustering results at the branching level. For example, *Anemone taipaiensis* and *Anemone trullifolia*, *Anemone narcissiflora* and *Anemone shikokiana*, *Anemone tomentosa* and *Anemone vitifolia*, etc. were clustered into one branch in both types of analyses. This suggests that the clustering analysis based on RSCU values can also reflect the affinities of buttercup plants to a certain extent, and the codon usage preferences of different plants are somehow related to their affinities. As for the deviation of the RSCU clustering results from the NJ-based phylogenetic tree, it may be due to the selection of only a single data, the RSCU value, and the combination of other data on codon preference may reduce the error.



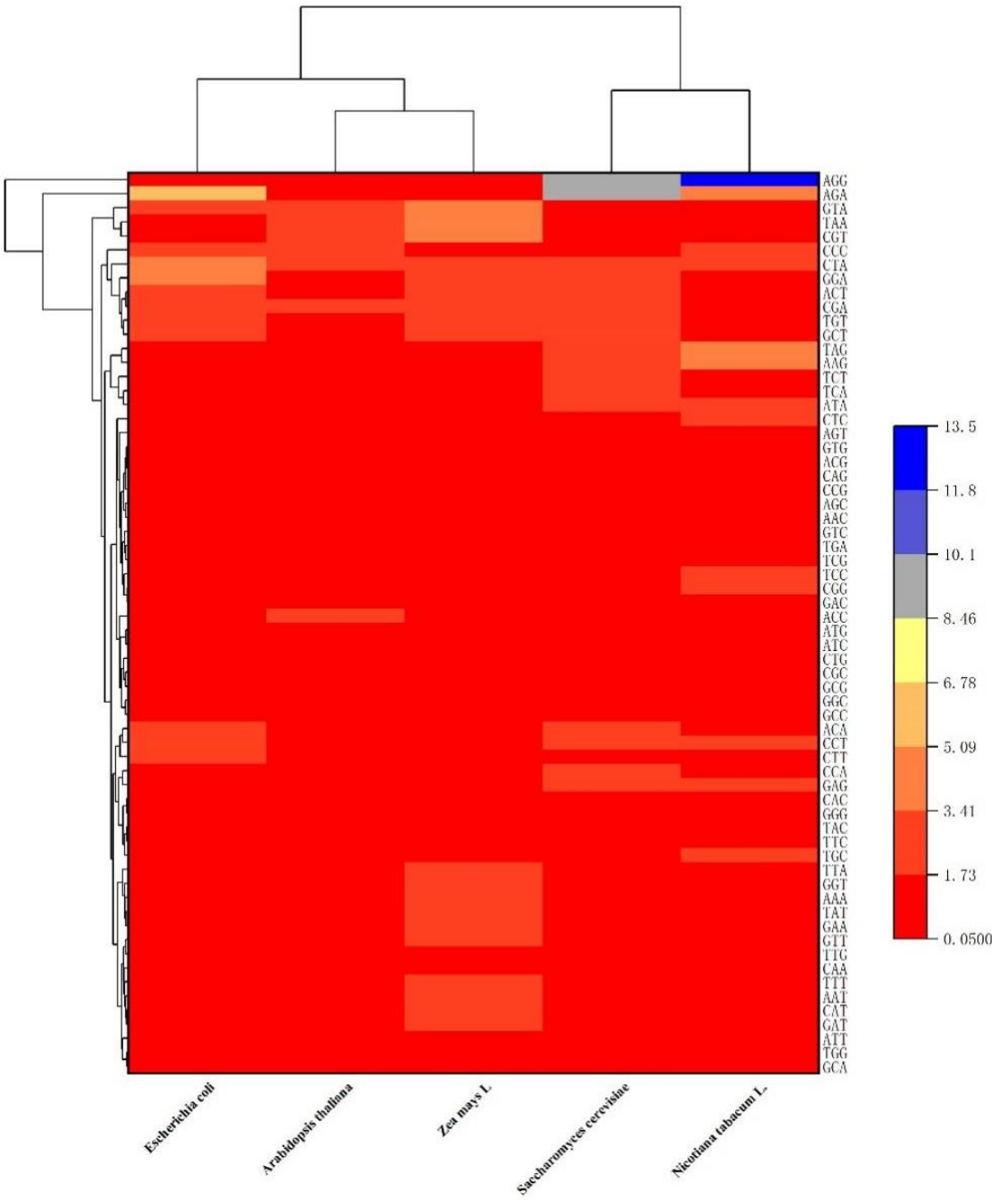
Figures 5. Construction of NJ tree based on CDS of *rbcL* genes. Test of phylogeny: bootstrap method, No. of bootstrap replications: 500; Substitution model: p-distance. Gaps/Missing data treatment: partial deletion, site coverage cutoff: 50%



Figures 6. Cluster analysis based on RSCU value construction. Method is averaging.

3.8. Comparative analysis of codon usage frequency

As can be seen in Figure 7, a comparison of the codon usage frequencies of the genome sequences of the 23 buttercup plant species and the model organisms shows that. Escherichia coli has 11 codons with a ratio greater than 2 and a maximum value of 6.527 (AGA); *Saccharomyces cerevisiae* has 12 codons with a ratio greater than 2 and a maximum value of 10 (AGG); indicating that compared to *Saccharomyces cerevisiae, Escherichia coli* is closer to the codon usage preference of plants in *Ranunculaceae*, so *Saccharomyces cerevisiae* can be used as an exogenous expression vector for the *rbcL* gene in *Ranunculaceae*; *Arabidopsis thaliana* had four codon ratios greater than 2, with a maximum value of 3.096 (CGT), and *Zea mays* had 14 codon ratios greater than 2 with a maximum value of 4.730 (GTA), and *Nicotiana tabacum* had 7 codon ratios greater than 2 with a maximum value of 13.454 (AGG). Therefore, *Arabidopsis thaliana* is more suitable as a transforming receptor for *rbcL* gene compared to *Zea mays* and *Nicotiana tabacum.*



Figures 7. Frequency analysis of codon usage in model organisms and Ranunculaceae plants. Among them, the data on the frequency of codon usage in Anemone are represented on the table in the form of proportions.

4. Conclusion

With the development of synthetic biology, the demand for mining natural biological resources has gradually increased [1]. My study analyzed the codon usage patterns and nucleotide composition of 22 *rbcL* gene of *Anemone*. I found that the codon usage patterns of *rbcL* of Anemone were mainly impacted by selection pressure. After RSCU analysis, the high-frequency codons often ended with A/T. The mean value of ENC value was 50.584, which indicated that the codon preference level of *rbcL* gene about *Anemone* was not high, and the correlation analysis showed that there was no significant correlation between CAI and ENC, which indicated that the expression level of the *rbcL* gene did not have any significant effect on the codon preference. In addition, I constructed the relationship between different *Anemone* with RSCU. After that, I found that the relationship between

different species, which were based on RSCU, was different from the sequence-based relationships. However, between some species, the genetic relationship was correctly interpreted according to the RSCU value. This result can be corroborated with the results of the phylogenetic tree based on nucleic acid sequences [3]. The above analyses indicate that RSCU-based phylogenetic tree construction has the feasibility that can be used for research and can be combined with existing phylogenetic tree construction strategies. In addition, I identified organisms that could serve as heterologous expression vectors for the *Anemone rbcL* gene: *Arabidopsis thaliana* and *Escherichia coli*. In summary, I studied the Codon usage bias of the *Anemone rbcL* gene. The results of codon usage bias provide theoretical basis for codon optimization of exogenous genes. This study adds to the understanding of the genetic evolution of *Anemone* and contributes to the discovery of *Anemone* resources for synthetic biology.

References

[1] Smith Reginald D. Enhanced effective codon numbers to understand codon usage bias. Bio Systems 220.(2022): 104734-104734.

[2] Xiong B, Wang T, Huang S, Liao L, Wang X, Deng H, Zhang M, He J, Sun G, He S, Wang Z. Analysis of Codon Usage Bias in Xyloglucan Endotransglycosylase (XET) Genes. Int J Mol Sci. 2023; 24(7): 6108.

[3] Hou Z, et al. Codon preference studies on *rbcL* genes of 11 species of Thalictrum (Thalictrum). Jiangsu Agricultural Science 51(3) (2023): 46-53.

[4] Li Q, Luo Y, Sha A, Xiao W, Xiong Z, Chen X, He J, Peng L, Zou L. Analysis of synonymous codon usage patterns in mitochondrial genomes of nine Amanita species. Front Microbiol. 2023; 14: 1134228.

[5] Gao M, Huo X, Lu L, Liu M, Zhang G. Analysis of codon usage patterns in Bupleurum falcatum chloroplast genome. Chin Herb Med. 2023; 15(2): 284-290..

[6] Sayers EW, Beck J, Bolton EE, Bourexis D, Brister JR, Canese K, Comeau DC, Funk K, Kim S, Klimke W, Marchler-Bauer A, Landrum M, Lathrop S, Lu Z, Madden TL, O'Leary N, Phan L, Rangwala SH, Schneider VA, Skripchenko Y, Wang J, Ye J, Trawick BW, Pruitt KD, Sherry ST. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2021; 49(D1): D10-D17.

[7] Ismail SNFB, Baharum SN, Fazry S, Low CF. Comparative genome analysis reveals a distinct influence of nucleotide composition on virus-host species-specific interaction of prawn-infecting nodavirus. J Fish Dis. 2019 Dec;42(12):1761-1772.

[8] Yu X , Liu J , Li H , et al. Comprehensive analysis of synonymous codon usage patterns and influencing factors of porcine epidemic diarrhea virus. Archives of Virology, 2020, 166(1): 1-9.

[9] Hao DC, Gu X, Xiao P. Anemone medicinal plants: ethnopharmacology, phytochemistry and biology. Acta Pharm Sin B. 2017 Mar;7(2):146-158.

[10] Kokate Prajakta PTechtmann Stephen M,and Werner Thomas. Codon usage bias and dinucleotide preference in 29 Drosophila species. G3: Genes, Genomes, Genetics 11.8(2021):

[11] Uddin A, Chakraborty S. Codon Usage Pattern of Genes Involved in Central Nervous System. Mol Neurobiol. 2019; 56(3): 1737-1748.

[12] Zheng G, et al. Codon usage preference analysis of the *rbcL* gene in 20 species of Chiquitaceae. Journal of Zhejiang Agriculture and Forestry University 38(3) (2021): 476-484.