Hu *et al. BMC Genomic Data* (2024) 25:80
https://doi.org/10.1186/s12863-024-01260-8

BMC Genomic Data

RESEARCH Open Access

Comparative analysis of codon usage bias in the chloroplast genomes of eighteen Ampelopsideae species (Vitaceae)

Qun Hu[1,2,3], Jiaqi Wu[1], Chengcheng Fan[1,2], Yongjian Luo[3], Jun Liu[3], Zhijun Deng[1,2*] and Qing Li[3*]

Abstract

Background The tribe Ampelopsideae plants are important garden plants with both medicinal and ornamental values. The study of codon usage bias (CUB) facilitates a deeper comprehension of the molecular genetic evolution of species and their adaptive strategies. The joint analysis of CUB in chloroplast genomes (cpDNA) offers valuable insights for in-depth research on molecular genetic evolution, biological resource conservation, and elite breeding within this plant family.

Results The base composition and codon usage preferences of the eighteen chloroplast genomes were highly similar, with the GC content of bases at all positions of their codons being less than 50%. This indicates that they preferred A/T bases. Their effective codon numbers were all in the range of 35–61, which indicates that the codon preferences of the chloroplast genomes of the 18 Ampelopsideae plants were relatively weak. A series of analyses indicated that the codon preference of the chloroplast genomes of the 18 Ampelopsideae plants was influenced by a combination of multiple factors, with natural selection being the primary influence. The clustering tree generated based on the relative usage of synonymous codons is consistent with some of the results obtained from the phylogenetic tree of chloroplast genomes, which indicates that the clustering tree based on the relative usage of synonymous codons can be an important supplement to the results of the sequence-based phylogenetic analysis. Eventually, 10 shared best codons were screened on the basis of the chloroplast genomes of 18 species.

Conclusion The codon preferences of the chloroplast genome in Ampelopsideae plants are relatively weak and are primarily influenced by natural selection. The codon composition of the chloroplast genomes of the eighteen Ampelopsideae plants and their usage preferences were sufficiently similar to demonstrate that the chloroplast genomes of Ampelopsideae plants are highly conserved. This study provides a scientific basis for the genetic evolution of chloroplast genes in Ampelopsideae species and their suitable strategies.

Keywords Ampelopsideae species, Chloroplast genome, Codon preference, Best codon, Cluster analysis

*Correspondence:
Zhijun Deng
dengzhijun@hbmzu.edu.cn
Qing Li
411066120@qq.com
[1]Hubei Key Laboratory of Biologic Resources Protection and Utilization, Hubei Minzu University, Enshi, Hubei 445000, China
[2]Research Center for Germplasm Engineering of Characteristic Plant Resources in Enshi Prefecture, Hubei Minzu University, Enshi, Hubei 445000, China
[3]Guangdong Key Laboratory for Crop Germplasm Resources Preservation and Utilization, Agro-biological Gene Research Center, Guangdong Academy of Agricultural Sciences, Guangzhou, Guangdong 510640, China

Background

Codons are the bridge between nucleic acids and proteins that carry and transmit instructions for genetic information. A total of 61 codons code for 20 amino acids, and different codons coding for the same amino acid are synonymous codons [1]. Synonymous codons exhibit markedly disparate probabilities of being utilized in translation. Moreover, species and genes exhibit varying degrees of predilection for the utilization of specific codons, which is referred to as codon usage bias (CUB) [2]. It has been reported that CUB is associated with a multitude of factors that have been implicated in the evolution of species. These include structural features of genes and proteins, gene expression, tRNA and protein abundance, the formation of mutation patterns, natural selection, and other related phenomena [3, 4]. Consequently, an investigation into the codon usage preferences of a given organism can facilitate the enhancement of the accuracy of transgenic loci and the optimization of the expression of exogenous genes [5]. Furthermore, such an inquiry can contribute to a deeper comprehension of the molecular genetic evolution of the organism in question and its adaptive strategies.

As semi-autonomous organelles, chloroplasts (cp.) are often postulated to have evolved from a symbiotic relationship between ancient bacteria and prokaryotic cells. They typically possess a distinct set of more conserved circular DNA [6]. Chloroplasts contain approximately 15% of total plant DNA and are often between 110 and 220 kb in length [7]. The small size and conserved structure of chloroplast genomes (cpDNAs) may be important reasons for the stable rate of cpDNA evolution and the slow rate of nucleotide turnover [8]. A significant number of genes were transferred from the chloroplast to the nucleus during the course of evolution. However, a series of proteins that are of critical importance for photosynthesis are still encoded by the cpDNA itself [9]. Consequently, the study of cpDNA can facilitate an understanding of its gene expression, an analysis of the evolutionary pattern and genetic relationship among plants. With the multifaceted exploration of cpDNA by scholars from various countries, cpDNA has the potential to contribute to a greater understanding of many fields, including the evolution of life genealogy, protein drug production, and chloroplast genetic engineering [10]. CpDNA-based codon preference studies have been conducted in a diverse range of plants, including members of the *Theaceae* [7], *Gynostemma* [11], *Oryza* [12], *Asteraceae* [13]. These studies have yielded valuable insights into the molecular genetic evolution and adaptive strategies of these species, which have significantly advanced the quality of breeding and exploitation of these species.

Ampelopsideae, as one of the small clades of *Vitaceae*, contains 4 genera and nearly 50 species of plants, widely distributed on all continents outside the boreal zone, mostly woody vines [14]. The type genus of this group is the genus *Ampelopsis*, which has simple, trifoliate, palmate or pinnately compound leaves, inflorescences that are usually hermaphroditic, opposite, dichasial cymes with 2-branched tendrils, and fruits of various colors, and is a monoecious climbing vine [15]. The plants of the genus Ampelopsis are ornamental plants with lush green leaves and fruits of different colors at different times of the year. In recent years, it has been reported that there are many medicinal species in the genus *Ampelopsis*, such as *Ampelopsis japonica* and *Ampelopsis delavayana*, which have antimicrobial activity, immunomodulation, and treatment of hypertension [16, 17]. The tender leaves and stems of *Nekemias grossedentata*, *Nekemias megalophylla* and *Nekemias cantonensis* are rich in dihydromyricetin, which has a special sweet taste and anti-glycemic and hepatoprotective effects, and has become the main source of vine tea, and has the opportunity to be used as a potential alternative health care for the prevention of high fat, high sugar and other related diseases [18, 19]. Today, the global habitat is deteriorating and the demand for vine tea is growing rapidly, leading to the shrinking of some of the wild resources of the Ampelopsideae family [20]. Currently, there are few studies on this family of species, and most of them are on pharmacological activities, dynamic evolutionary patterns of near-origin species, and systematic classification [14, 21–23]. However, little is known about its molecular genetic aspects, which is not conducive to the subsequent conservation and development of the species in this lineage. Therefore, in the present study, we combined 18 cpDNAs to perform a comparative codon preference analysis with the aim of providing important references for in-depth studies on molecular genetic evolution, conservation of biological resources, and selection of superior species of the Ampelopsideae family.

Results

Analysis of bias in codon base composition

As shown in Table 1, the cpDNAs of each species have their respective corresponding numbers (A-R), which are presented in the analyzed figures below. It is evident that the nucleotide compositions of these 18 cpDNAs are highly similar. Their GC1, GC2 and GC3 contents ranged from 45.56% (*R. digitata*) ~ 46.44% (*A. aconitifolia var. palmiloba*), 37.76% (*N. rubifolia*) ~ 38.67% (*A. aconitifolia* var. *palmiloba*) and 29.87% (*R. digitata*) ~ 30.52% (*A. aconitifolia* var. *palmiloba*), and the GCall content ranged from 37.74% (*R. digitata*) ~ 38.55% (*A. aconitifolia* var. *palmiloba*). The codon contents of GC1, GC2 and GC3 were all less than 50% and showed a decreasing pattern, i.e. GC1 > GC2 > GC3. It was obvious that the codons of the 18 cpDNA coding sequences preferred A/T

Table 1 Fundamental parameters of CUB of cpDNA in Ampelopsideae

Numbering	Species	GCall%	GC1%	GC2%	GC3%	ENC
A	*Ampelopsis aconitifolia* var. *palmiloba*	38.55	46.44	38.67	30.52	47.178
B	*A. cordata*	37.88	45.62	38.01	30.01	47.013
C	*A. delavayana*	38.01	45.91	38.06	30.06	47.160
D	*A. glandulosa* var. *brevipedunculata*	37.89	45.62	37.94	30.11	47.191
E	*A. heterophylla*	37.80	45.59	37.83	29.96	47.335
F	*A. heterophylla* var. *hancei*	37.80	45.59	37.85	29.96	47.361
G	*A. humulifolia*	37.94	45.73	37.97	30.12	47.194
H	*A. japonica*	37.96	45.78	37.96	30.13	47.165
I	*A. mollifolia*	37.84	45.67	37.94	29.90	47.228
J	*A. tomentosa*	37.83	45.67	37.93	29.91	47.220
K	*Nekemias arborea*	37.84	45.66	37.83	30.05	46.942
L	*N. cantoniensis*	38.05	45.91	38.06	30.18	47.255
M	*N. chaffanjonii*	37.86	45.7	37.86	30.01	47.220
N	*N. grossedentata*	37.88	45.57	37.84	30.22	47.573
O	*N. hypoglauca*	37.86	45.7	37.88	30.01	47.220
P	*N. megalophylla*	38.00	45.87	38.02	30.12	47.150
Q	*N. rubifolia*	37.81	45.63	37.76	30.06	47.402
R	*Rhoicissus digitata*	37.74	45.56	37.80	29.87	46.853

bases and ended with A/T bases. The differences in the content of GC bases at the three positions in the codons were all within 1%, indicating that the CUBs of these 18 cpDNAs were also very similar.

Analysis of 18 neutrality plots

From Fig. 1, it can be seen that the genes of the 18 cpDNAs are all above the straight line y=x, and the codon GC3 are all between 17.70% and 36.98%, and GC12 are all between 33.05% and 56.48%, indicating that the bases of codon position 3 are more skewed toward A/T. The slopes of the regression fitting curves of each neutral analysis range from −0.0056 to 0.105, and the R2 ranges from 0 ~ 0.0076, indicating that the correlation between GC12 and GC3 is not significant, and that the base profiles at each site of the codons of the 18 cpDNAs are more diverse, with CUB more likely to have been affected by natural selection. The slope of the regression curve for *A. heterophylla* (-0.0019) was closest to 0, and thus the CUB of its cpDNA was most affected by natural selection, while the slope of the regression curve for *N. megalophylla* (0.105) was most different from 0, and thus the CUB of its cpDNA was relatively least affected by natural selection.

Analysis of 18 ENC-GC3s plots

The ENC-plots of the 18 cpDNAs were extremely similar (Fig. 2), with some of the gene scatters converging to the ideal ENC curve, indicating that their ENCs are similar to the ideal values and their CUBs are most affected by the base mutation factor; and some of the gene scatters deviating from the ideal curve, showing that the actual ENC values are quite different from the ideal values, also indicating that the base mutation is not the main factor affecting their CUBs. This also implies that the CUB of the 18 cpDNAs may be disturbed by multiple factors such as natural selection and base mutation. The ENC values of all 18 cpDNAs in this study ranged from 35.55 to 57.53, and their mean value was 47.194 (Supplementary Table 1). All ENCs ranged from 35 to 61, indicating that the CUB of these 18 cpDNAs was weak.

Analysis of 18 PR2-plots

PR2-plot analysis is often used to visualize the distribution of bases at codon 3 to explore the multiple factors affecting CUB. Combined with the PR2 comparative analysis of the 18 cpDNAs, it can be seen that the scatters are unevenly distributed in the four regions in Fig. 3. From the top and bottom distributions, it can be seen that a large portion of the scatters are distributed in the lower half of the region (quadrants 3 and 4), suggesting that the selection of A/T bases in position 3 tends to be more T. From the left and right distributions, it can be seen that most of the scatters are distributed in the right half of the region (quadrants 1 and 4), suggesting that the selection of G/C bases in position 3 tends to be more G. Combined with the comparison in the four quadrants, it can be seen that the scatters distributed in the fourth quadrant tend to be more G. The comparison of the four quadrants shows that the largest number of scatters are distributed in quadrant 4, indicating that the selection of codon 3 bases tends to be more in favor of G/T, which also suggests that natural selection is the dominant factor leading to CUB.

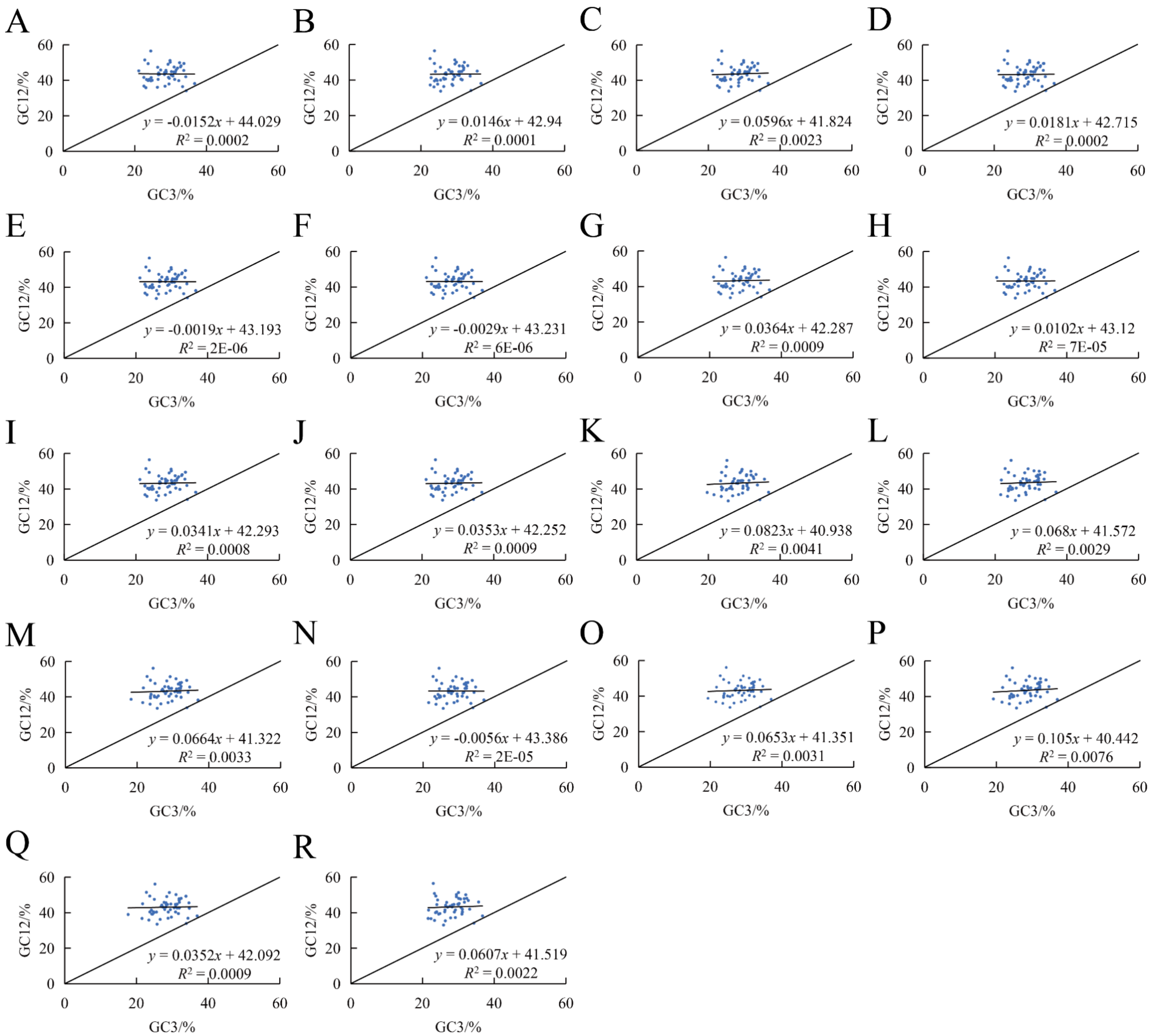

Fig. 1 Neutrality plot analysis of 18 cpDNAs in Ampelopsideae

Analysis of 18 COA-plots

To further explore the codon bias of these 18 species cpDNAs, we performed COA mapping based on the RSCU values of the CDSs (Fig. 4), and the dots representing the different genes were segregated from each other in all COA-plots. In the 18 cpDNAs, the first four principal axes collectively account for 32.28–35.45% of the variation in synonymous codons. Among them, Axis 1 contributes the largest proportion of the total variation (9.38–10.76%), followed by Axis 2 (8.09–9.13%) (Supplementary Table 2). The major factor axes accounted for a decreasing proportion of the total variation as the axis order increased, indicating that the CUB was the result of multiple factors working together. Most of the genes in the COA plots of the 18 species were centered and centrally distributed, indicating that most of the genes had a similar CUB.

From the analysis of the correlation between the first dimension axes and the CAIs, CBI, Fop, ENC, and GC3s (Table 2) shows that the CBI and Fop values of the 18 species are significantly or very significantly positively or negatively correlated with the first dimension axis, and the CAI and GC3s are significantly or very significantly positively or negatively correlated with the first dimension axis, and the CAI and GC3s values of some species are significantly or very significantly positively or negatively correlated with the first dimension axis, suggesting that the formation of the codon usage patterns of the 18 cpDNAs is a complex process influenced by a combination of several factors.

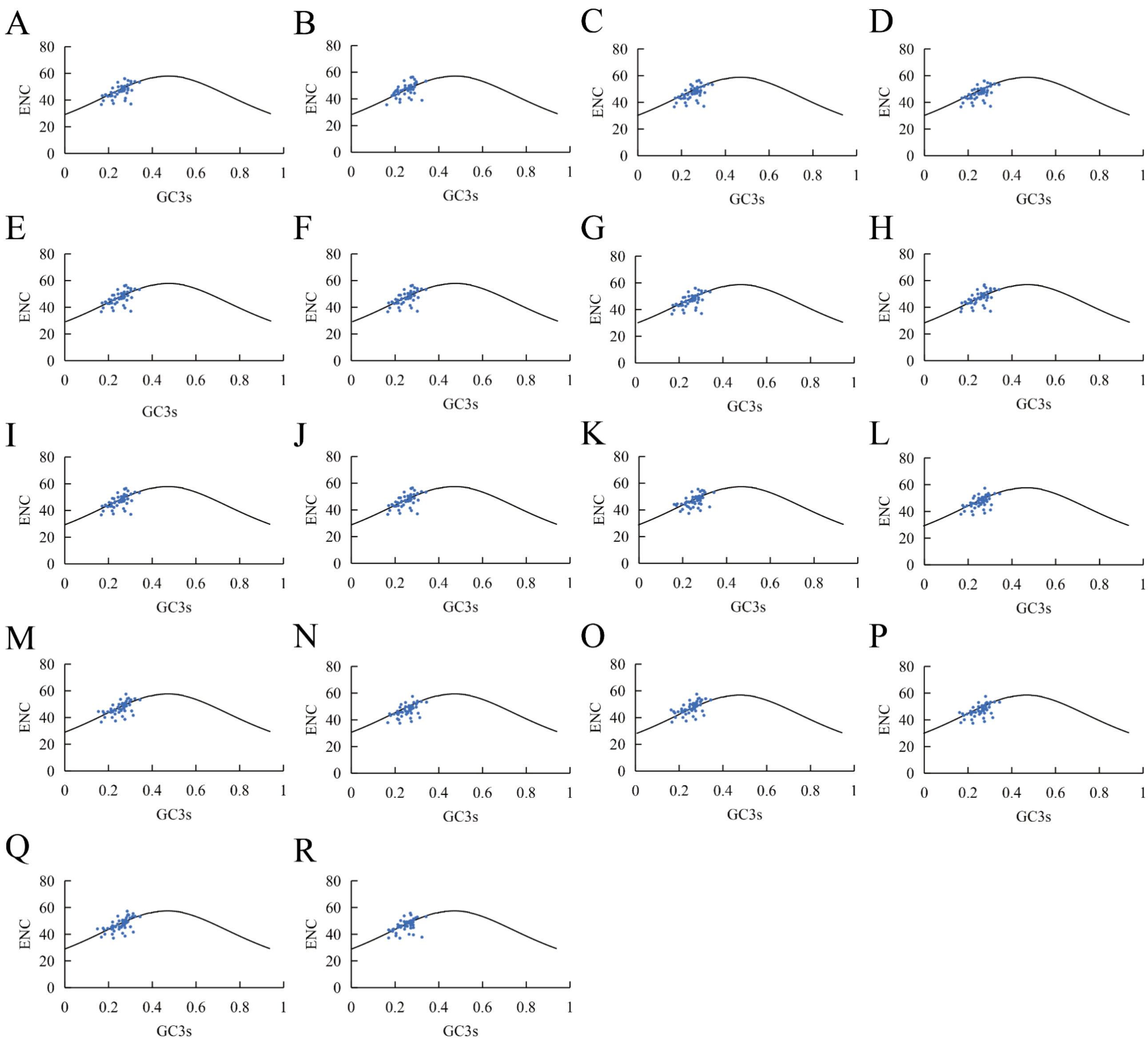

Fig. 2 ENC-plots analysis of 18 cpDNAs in Ampelopsideae

RSCU analysis and phylogeny of 18 cpDNAs

RSCU, as the ratio of the actual codon usage frequency to the ideal codon usage frequency, is an important index to measure CUB. 18 cpDNAs had RSCU values ranging from 0.322 to 1.866, among which there were 29 high-frequency codons, 28 of which ended in U/A (96.55%) and 1 in G, indicating that the probability of preferred codons ending in U/A is much higher than that of codons ending in G/C (Fig. 5). Among the 18 synonymous codons of the cpDNA, the codon with the highest RSCU value was AGA encoding arginine (Arg), followed by GCU encoding alanine (Ala).

A phylogenetic tree based on the CDSs of 18 cpDNAs (Fig. 6) shows that the phylogenetic tree divides the 18 species into three branches: seven plants of the genus *Nekemias* are classified in one group; ten plants of the genus Ampelopsis were classified in one group; *Rhoicissus digitata* of the genus *Rhoicissus* was a separate category. All branches in this phylogenetic tree have bootstrap values greater than 70, indicating a high level of confidence in the evolutionary analysis results of this study. Compared with the former, the RSCU-based clustering tree showed significant differences in species associations. However, it also clearly illustrates the close relationship among *Ampelopsis japonica*, *Ampelopsis humulifolia*, and *Ampelopsis glandulosa* var. *brevipedunculata* within the genus *Ampelopsis*, as well as the intimate relationships among *Nekemias chaffanjonii*, *Nekemias hypoglauca*, and *Nekemias rubifolia* within the genus *Nekemias*.

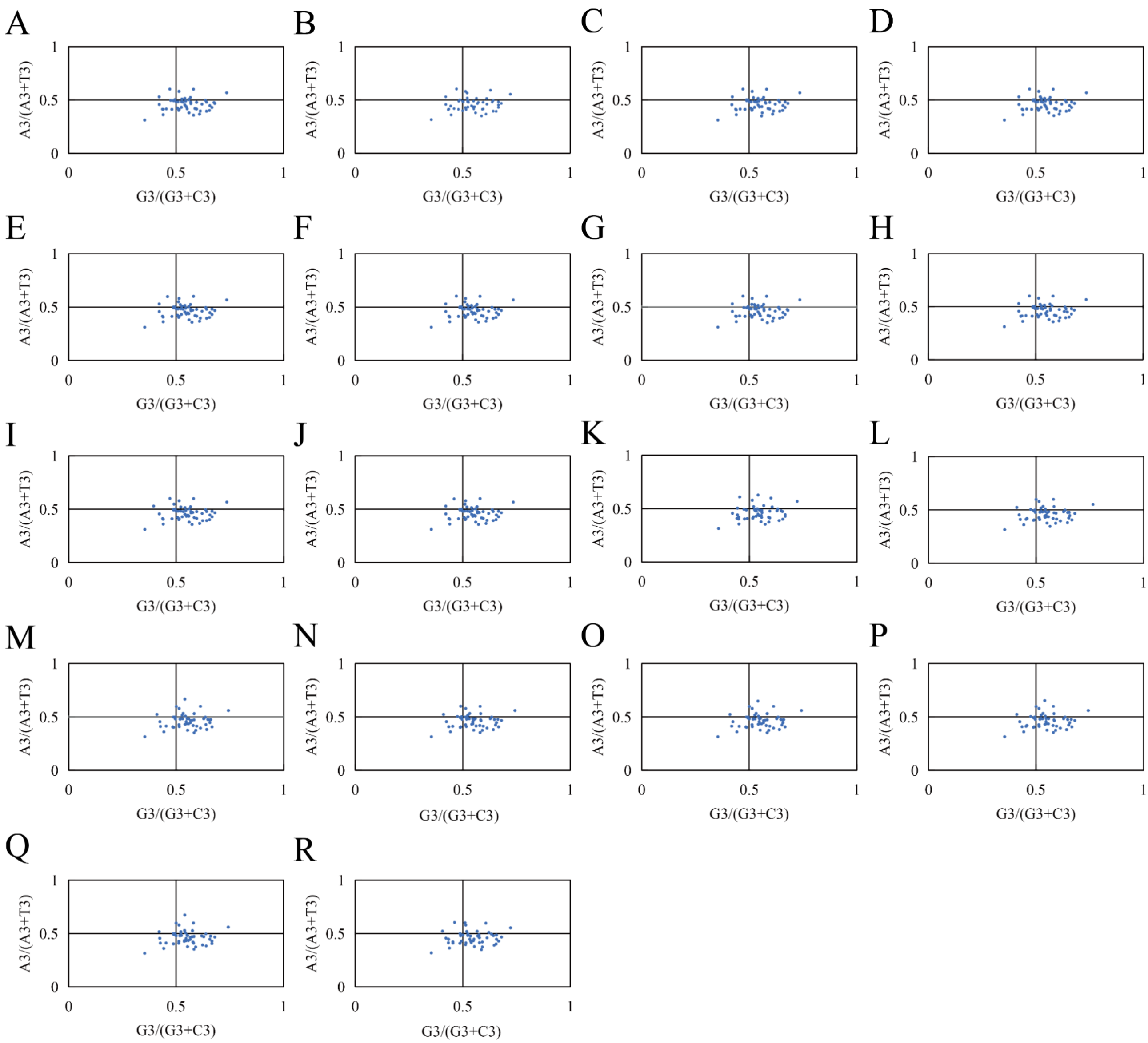

Fig. 3 PR2-plots analysis of 18 cpDNAs in Ampelopsideae

Comparative analysis of the best codons of 18 cpDNAs

Twenty-six co-occurring high expression codons and 10 co-occurring best codons (GCU, UGU, GAA, AUU, AAA, UUG, CCU, ACU, GUA, and GUU) were screened in the 18 cpDNAs, of which nearly 90% of the high expression codons ended in U/A (Fig. 7). In addition, there were some deviations in the corresponding high expression codons and best codons of the 18 different species, in contrast to the other species, GGC of *N. megalophylla* was the high expression codon, GGU and CGU of *N. cantoniensis* and CUU of *A. mollifolia* were not the high expression codon and best codon, and the high expression codons and best codons of *N. grossedentata*, *N. hypoglauca*, *N. megalophylla*, *N. rubifolia*, and *R. digitata* of GCA were not optimal codons, *A. glandulosa* var. *brevipedunculata*, *A. heterophylla*, *A. heterophylla* var. *hancei*, *A. humulifolia*, and *A. japonica* had non-optimal codons for CAA, and *A. aconitifolia* var. *palmiloba* and *A. cordata* had non-optimal codons for CGA. This diversity suggests that some discrimination between species can be made on the basis of highly expressed or optimal codons.

Discussion

Genetic codons, as the core elements linking nucleic acids and proteins, play an important role in the exchange of genetic information in organisms [24]. And genetic mutations drive the evolutionary process through codon changes and are expressed in the form of proteins [25]. Consequently, an investigation of CUB can furnish

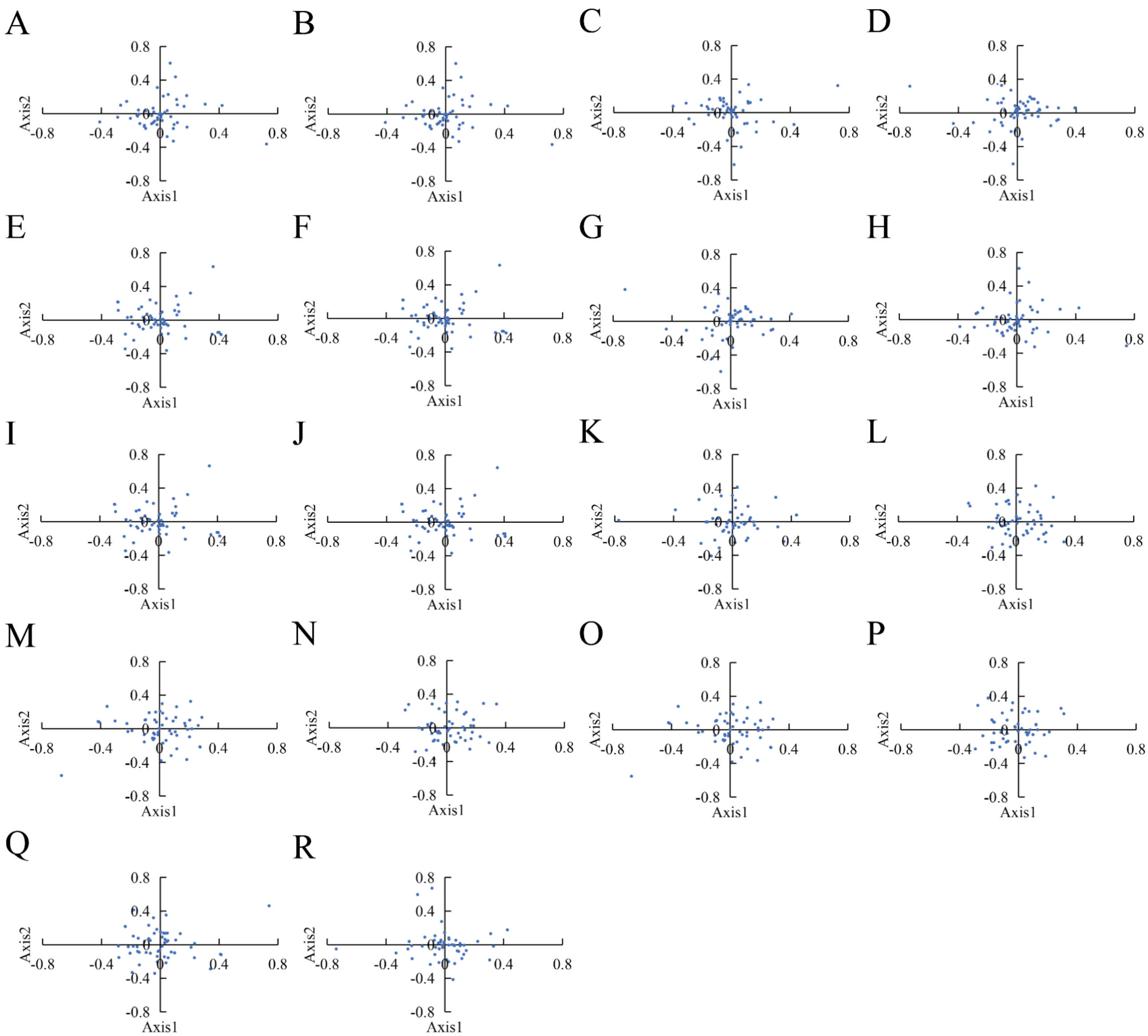

Fig. 4 COA-plots analysis of 18 cpDNAs in Ampelopsideae

dependable data for the study of protein expression and its associated functions [26]. In the process of studying the genetic evolution of species, we often employ nuclear genome sequencing and organelle genome sequencing. However, compared to nuclear genome sequencing analysis, organelle genomes possess a number of advantages, including a relatively smaller number of sequences, a lower sequence complexity, a single genetic origin, a very low frequency of recombination, and greater maneuverability [27–29]. It is now evident that organelle genomes play a pivotal role in the study of the origin and evolution of species. However, their synonymous codon usage preference differs from that of nuclear genes in terms of the rate and pattern of evolution. This is evidenced by the fact that it is not only subjected to genetic variation and the result of natural selection, but also related to multiple factors, such as the base composition of genes, the length of genes, the level of gene expression, the abundance of tRNAs, the hydrophobicity of amino acids, and the aromaticity of genes [30]. Our multiple analyses (codon usage index, Neutrality-plot analysis, ENC-plot analysis, PR2-plot analysis, best codon analysis, COA-plot analysis, and evaluation of RSCUs and phylogenetic analyses) contributed to our better understanding of the genetic structure and evolutionary trends of the species in this strain.

Chloroplasts, as a common organelle structure in higher plants, have their own peculiarities in the evolutionary history of plants, with relatively conserved gene structural domains and the ability to be inherited through

Table 2 Correlation analysis of codon usage index of 18 cpDNAs in Ampelopsideae

Species	CAI	CBI	Fop	ENC	GC3s
A. aconitifolia var. *palmiloba*	-0.171	-0.419**	-0.448**	-0.162	-0.524**
A. cordata	-0.019	-0.288*	-0.328*	0	-0.403**
A. delavayana	-0.197	-0.453**	-0.478**	-0.1	-0.49**
A. glandulosa var. *brevipedunculata*	0.171	0.416**	0.448**	0.138	0.501**
A. heterophylla	-0.264*	-0.48**	-0.42**	0.131	-0.029
A. heterophylla var. *hancei*	-0.255	-0.473**	-0.414**	0.126	-0.035
A. humulifolia	0.161	0.412**	0.445**	0.178	0.527**
A. japonica	-0.172	-0.421**	-0.449**	-0.119	-0.493**
A. mollifolia	-0.266*	-0.473**	-0.411**	0.14	-0.011
A. tomentosa	-0.269*	-0.482**	-0.42**	0.143	-0.019
N. arborea	0.119	0.329*	0.373**	-0.024	0.406**
N. cantoniensis	0.242	0.421**	0.432**	-0.148	0.177
N. chaffanjonii	0.249	0.459**	0.403**	-0.099	-0.033
N. grossedentata	0.176	0.346**	0.36**	-0.22	0.116
N. hypoglauca	0.256	0.475**	0.419**	-0.099	-0.025
N. megalophylla	-0.191	-0.39**	-0.401**	0.188	-0.122
N. rubifolia	-0.221	-0.466**	-0.42**	0.114	-0.009
R. digitata	0.099	0.381**	0.412**	-0.069	0.375**

"**" Highly significant correlation (P<0.01); "*" Significant correlation (P<0.05)。

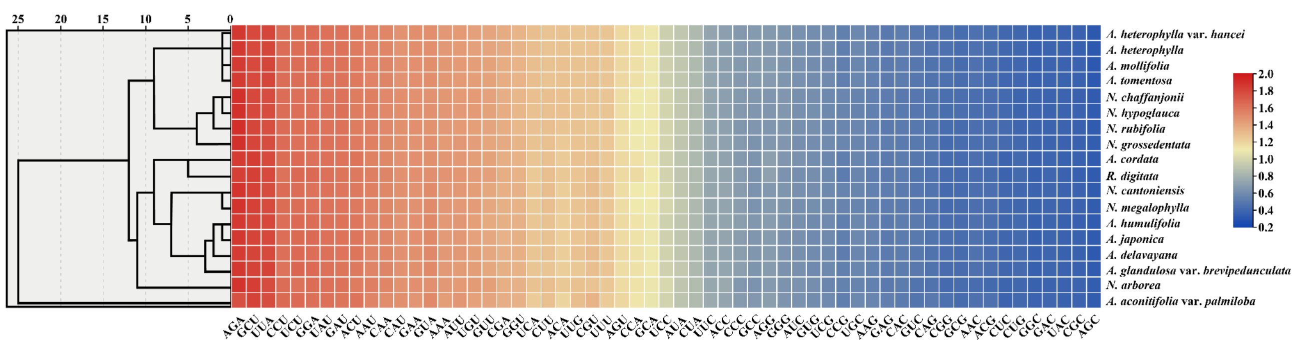

Fig. 5 RSCU clustering tree and heat map of cp. genes in Ampelopsideae

autogenous inheritance [31]. Among the codons of chloroplast genes, the GC3 content is an important indicator in the analysis of their preference, and in general, the smaller the GC3, the more the codon usage preference is influenced by natural selection, and if there is no significant correlation between GC12 and GC3, the more it is influenced by natural selection [27]. The codon bases of the 18 cpDNAs under investigation exhibit a clear preference for either A or U. Furthermore, they exhibit a high degree of similarity in base content and CUB. Moreover, each neutral analysis demonstrated that the correlation between GC12 and GC3 was not statistically significant. This result is analogous to that of *Nymphaea*, *Juglandaceae*, *Davidia*, and *Gynostemma*, indicating that the CUBs of the 18 cpDNAs are more susceptible to natural selection, and that the CUBs of the cpDNAs of most plants are more conserved and highly similar during the evolutionary process [11, 27, 32, 33]. The ENC-plot analysis indicated that the CUB of the cpDNAs of the 18 Ampelopsideae species may be influenced by a multitude of factors, including natural selection and base mutations. To further explore the main factors affecting CUB, this study found that most of the scatters were concentrated in the fourth quadrant by PR2-plot bias analysis. This indicated that the base selection at codon 3 was more inclined to G/T, thus suggesting that natural selection was one of the dominant factors leading to CUB. RSCU is a frequently employed term to describe the strength of CUB. According to the COA analysis of RSCU, the first dimension axis was found to be significantly correlated with CAI, CBI, Fop, ENC, and GC3s. The results of multiple analyses indicated that the CUB of the cpDNAs of the 18 Ampelopsideae plants was primarily influenced by natural selection, in addition to multiple factors such as base mutation and gene expression frequency. This finding was consistent with the results of previous studies on dicotyledonous plants in *Theaceae*, *Euphorbiaceae*, and *Juglandaceae* [7, 32, 34].

The clustering tree constructed based on RSCU values exhibited significant divergence from the evolutionary

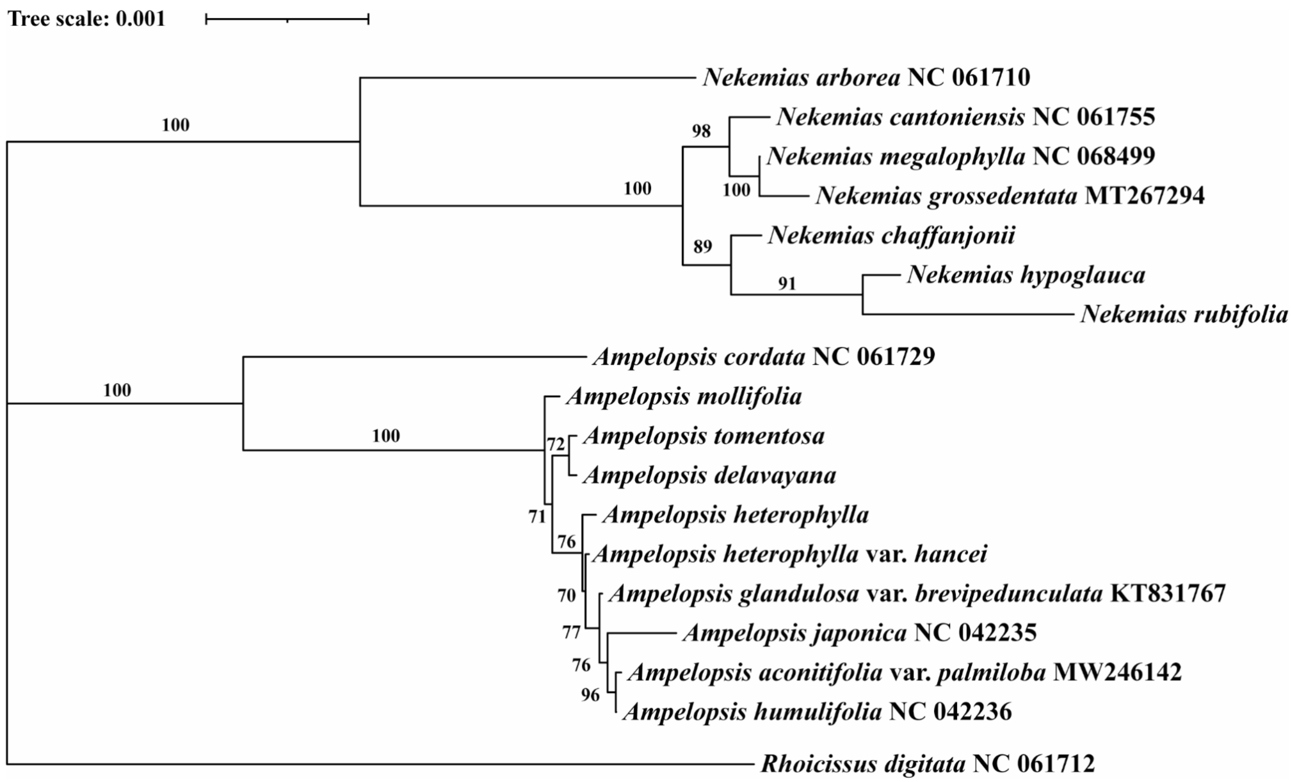

Fig. 6 Phylogenetic tree based on the CDSs of 18 cpDNAs. The numerical values on each branch in the phylogenetic tree represent their bootstrap values, where a higher bootstrap value indicates greater confidence in that particular branch's result

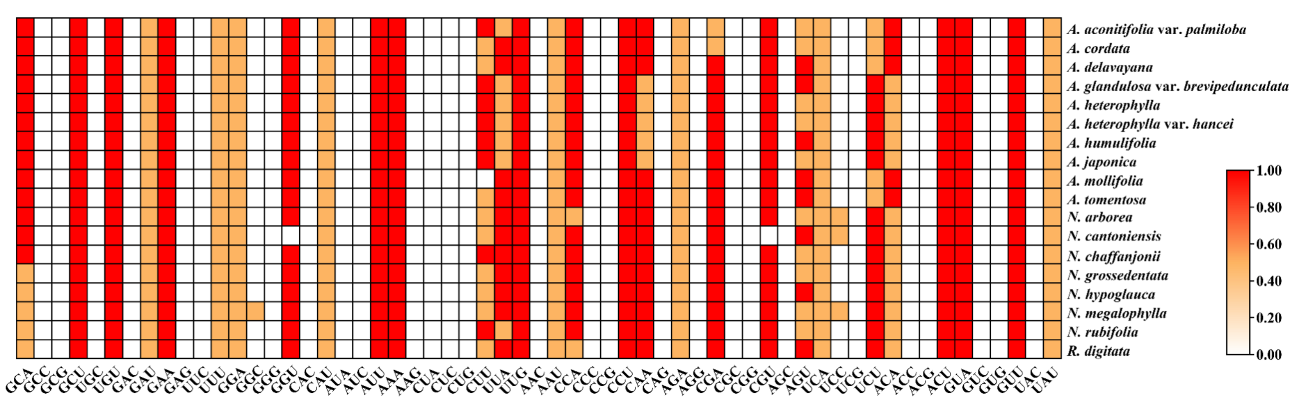

Fig. 7 Heatmap of bast codons in 18 cpDNAs

tree constructed based on CDSs. However, some species relationships exhibited consistent branching patterns in both trees. It appears that the phylogenetic prediction based on the RSCU values of cpDNA codons is subject to certain limitations, which may be attributable to the considerable geographic differences among the sampled species. It is evident that RSCU-based clustering trees can serve as a valuable addition to sequence-based phylogenetic predictions [30]. The evaluation of the 18 RSCUs revealed that their RSCU values were strikingly similar, thereby indicating that the cpDNA codon composition of Ampelopsideae species is indeed highly conserved and approximate. The selection of synonymous codons not only impacts the coding of protein sequences but also the regulation of gene expression [35]. In the field of biology, correlations between codon usage and expression levels are often attributed to evolutionary or physiological causal links. Optimal codon usage is associated with faster translation rates, which suggests that some proteins will have higher expression levels as a result of codon usage [35]. In this study, 26 shared high-frequency codons and 10 shared optimal codons were identified. Of these, 90% of the optimal codons terminated in U/A. It is noteworthy that there are discrepancies in the high-expression codons and optimal codons of closely related species. These differences may be leveraged in future studies to categorize species based on their codon usage patterns. The high-frequency codons identified based on the total RSCU value also exhibited a tendency to end in U/A. This indicated that they were more likely to select U/A at position 3, consistent with model plants such as rice, poplar, and Arabidopsis thaliana [36–38]. These optimal codons can, on the one hand, enhance the yield and activity of proteins involved in photosynthesis by optimizing the expression of cp. genes. On the other hand, they can be employed as a breakthrough to increase the production of biopharmaceuticals [39, 40].

Conclusions

The cpDNAs of 18 species of the tribe Ampelopsideae are weak in CUB, which is affected by multiple factors such as natural selection, base mutation and gene expression frequency, among which natural selection is the main factor. The 18 cpDNAs were screened for 10 common best codons, which can provide some reference for their biopharmaceuticals. Their RSCU clustering tree can be used to supplement the phylogenetic results. The codon compositions and CUBs of the 18 cpDNAs are similar enough to show that the cpDNAs of Ampelopsideae plants are highly conserved and similar, and the comparative analysis of these 18 cpDNA codons can provide a certain scientific research basis for the CUBs of cpDNAs of the tribe Ampelopsideae plants as a whole.

Materials & methods

The source and processing of cpDNA sequences

The cpDNA sequences required for this study were obtained from the NCBI database (https://www.ncbi.nlm.nih.gov/nuccore/) and from our group (not submitted to publicly available databases, https://github.com/mrprinceq/Supplementary-data-1-8-eight-cpDNAs-). To ensure the accuracy of subsequent analysis results, filter the coding sequences (CDS) of 18 cpDNA species as follows: retain sequences composed solely of A, T, C, and G bases, with a length that is a multiple of 3 and not less than 300 bp, starting with a start codon, ending with a stop codon, and containing no internal stop codons [34]. As shown in Table 3, the cpDNA sizes of 18 species range from 160,600 bp (*R. digitata*) to 163,016 bp (*N. arborea*), with the number of CDS ranging from 80 (*R. digitata*) to 92 (*N. chaffanjonii*, *N. hypoglauca*, *N. rubifolia*). After filtering, the retained number of CDS ranges from 54 to 59.

Composition and codon index analysis

In order to obtain the details of the codon base contents of the 18 cpDNAs, the retained CDS files after filtering were uploaded to the online software (http://emboss.toulouse.inra.fr/cgi-bin/emboss) on the EMBOSS website for computational analysis [41]. CodonW is a program that simplifies codon usage and calculates CUBs. Use CodonW 1.42 to calculate the relative usage rate of synonymous codons (RSCU), effective codon count (ENC), and codon adaptation index (CAI), among others [42]. RSCU is used to measure the relative overall usage of a codon compared to its encoded amino acid. An RSCU value of 1 serves as a threshold: RSCU > 1 indicates

Table 3 Basic information of chloroplast genome in Ampelopsideae

Data source	Species	Accession no	Length	CDS number (before processing)	CDS number (after processing)
NCBI	*Ampelopsis humulifolia*	NC_042236.1	161,724 bp	87	56
	A. aconitifolia var. *palmiloba*	MW246142.1	162,493 bp	84	54
	A. cordata	NC_061729.1	161,945 bp	82	54
	A. glandulosa var. *brevipedunculata*	KT831767.1	161,090 bp	87	57
	A. japonica	NC_042235.1	161,430 bp	87	55
	Nekemias arborea	NC_061710.1	163,016 bp	82	54
	N. cantoniensis	NC_061755.1	162,655 bp	85	55
	N. grossedentata	MT267294.1	162,147 bp	88	55
	N. megalophylla	NC_068499.1	161,981 bp	85	55
	Rhoicissus digitata	NC_061712.1	160,600 bp	80	54
Subject assembly	*A. delavayana*	-	162,497 bp	85	55
	A. heterophylla	-	162,470 bp	91	59
	A. heterophylla var. *hancei*	-	162,490 bp	91	59
	A. mollifolia	-	162,539 bp	90	58
	A. tomentosa	-	162,468 bp	90	58
	N. chaffanjonii	-	163,013 bp	92	58
	N. hypoglauca	-	162,664 bp	92	58
	N. rubifolia	-	162,813 bp	92	59

strong codon usage bias (CUB), whereas RSCU<1 indicates weaker CUB [43]. ENC is used to assess the statistical indices of Codon Usage Bias (CUB) in genes, generally ranging between 20 and 61. A threshold of ENC=35 serves as a dividing point: an ENC of 35 or less indicates significant CUB, with lower values suggesting stronger CUB within the gene [44].

Neutral plotting analysis

Neutral plot is one of the methods used to intuitively display codon preference indices of cpDNA genes. It plots the GC content at the third codon position (G3) on the horizontal axis against the mean GC content at the first and second codon positions (G12) on the vertical axis. Scatter plots depicting the correlation between codon base composition are created for 18 cpDNA genes, aimed at analyzing determinants of Codon Usage Bias (CUB) [45, 46]. If the regression coefficient fitted by the scatter plot approaches 0, the correlation between GC3 and GC12 becomes less significant, suggesting that Codon Usage Bias (CUB) may be dominated by natural selection. Conversely, if the regression coefficient approaches 1, it indicates a significant correlation between the two, suggesting that mutation pressure determines CUB [47].

ENC-GC3s plotting analysis

To analyze the relationship between Codon Usage Bias (CUB) and base composition across 18 cpDNAs, scatter plots were created with ENC as the y-axis and GC3s as the x-axis for each gene. Additionally, a standard curve based on the following formula was added to the plots [48]. The standard curve represents the ideal relationship between ENC and GC3 under conditions of no selection pressure. When the scatter points on the graph closely align with this curve, it indicates that Codon Usage Bias (CUB) for the corresponding gene is solely determined by mutation. Conversely, deviation from the curve suggests the presence of other factors influencing CUB [49].

$$ENC = 2 + GC3s + \frac{29}{\left[GC3s^2 + (1 - GC3s)^2\right]}$$

PR2 bias plotting analysis

To analyze codon bias equilibrium across 18 cpDNA genes, scatter plots were created with the proportion of bases A and AT at the third codon position on the y-axis, and the proportion of bases G and CG on the x-axis [50]. x=0.5 and y=0.5 represent G=C and A=T, respectively, and the center point of the scatterplot (0.5, 0.5) indicates that the codon has no usage preference.

Correspondence analysis

Correspondence analysis (COA), as a method applicable to multivariate statistics, is commonly used to identify codon usage patterns. Using the COA feature in CodonW software, each CDS from the 18 cpDNAs is represented in a 59-dimensional vector space (excluding UGG, AUG, and the three termination codons). This analysis examines the major trends of variation in codon usage (CU) within the CDSs, ultimately aligning codons along axes based on their RSCU values. This approach aims to provide a clearer and more direct visualization of the codon usage variation trends across all coding sequences (CDS) of each species [51]. To better understand the factors influencing Codon Usage Bias (CUB) across the 18 cpDNAs, correlation analyses were conducted between the first principal component axis and CAI, CBI, Fop, ENC, and GC3s. Data organization and statistical analyses were performed using WPS and SPSS 26 software.

Evaluation of RSCU and phylogenetic analysis

The RSCU values (removing UGG, AUG and three stop codons) of all genes of the 18 cpDNAs were organized in one dataset and heatmapped on TBtools software. In order to reveal the phylogenetic relationships of the 18 species, we analyzed two aspects: (1) making a clustering tree based on the RSCU values of the codons of the 18 cpDNAs on SPSS software; (2) performing multiple sequence comparison based on public CDSs of 18 cpDNAs in MAFFT software [52]. The sequence comparison results were then imported into the IQ-TREE (v.1.6.8) software, and a phylogenetic tree was constructed using the best-fit model (TVM+F+I chosen according to BIC) and the maximum likelihood method with 1000 replications [53].

Best codon analysis

Based on the RSCU data can be parsed out the codons with high frequency of use of RSCU values, the ENCs of each CDS of the 18 cpDNAs were sorted, and the genes in the first and last 1/10 of the ENCs were selected to construct the high- and low-expression databases, respectively, and the difference between the two RSCUs was used as the ΔRSCU [54]. Codons with RSCU greater than 1 were classified as high usage frequency codons, while codons with ΔRSCU not less than 0.08 in addition to the previous conditions were recognized as the best codons.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12863-024-01260-8.

Supplementary Material 1

Supplementary Material 2

Acknowledgements
The authors acknowledge the availability of cp. genome data in public databases, which facilitated this study.

Author contributions
Q.L. (Qing Li) and Z.D. (Zhijun Deng) planned and supervised the project, conceived and designed the experiments, and participated in obtaining funding; Q.H. (Qun Hu) and J.W. (Jiaqi Wu) wrote the manuscript; C.F. (Chengcheng Fan), Y.L. (Yongjian Luo) and J.L. (Jun Liu) performed the data organization and formal analysis. All authors have read and agreed to the published version of the manuscript.

Funding
This work was supported by the Li Qing Youth Mentorship Project (0002002094) and the 2023 Provincial Rural Revitalization Strategy Special Funds Seed Industry Revitalization Project, Operation and Maintenance of Perennial Special Cash Crop Resource Nurseries in Guangdong Province (2023-NBH-00-017).

Data availability
The dataset used for the analyses in this study includes ten chloroplast genomes sourced from publicly available databases and eight chloroplast genomes assembled by our research group in previous stages of this project. The ten chloroplast genomes sourced from public databases are available in the NCBI, Ampelopsis humulifolia (NC_042236.1, https://www.ncbi.nlm.nih.gov/nuccore/NC_042236.1/), Ampelopsis aconitifolia var. palmiloba(MW246142.1, https://www.ncbi.nlm.nih.gov/nuccore/MW246142.1), Ampelopsis cordata (NC_061729.1, https://www.ncbi.nlm.nih.gov/nuccore/NC_061729.1), Ampelopsis glandulosa var. brevipedunculata (KT831767.1, https://www.ncbi.nlm.nih.gov/nuccore/KT831767.1), Ampelopsis japonica (NC_042235.1, https://www.ncbi.nlm.nih.gov/nuccore/NC_042235.1), Nekemias arborea (NC_061710.1, https://www.ncbi.nlm.nih.gov/nuccore/NC_061710.1), Nekemias cantoniensis (NC_061755.1, https://www.ncbi.nlm.nih.gov/nuccore/NC_061755.1), Nekemias grossedentata (MT267294.1, https://www.ncbi.nlm.nih.gov/nuccore/MT267294.1), Nekemias megalophylla (NC_068499.1, https://www.ncbi.nlm.nih.gov/nuccore/NC_068499.1), and Rhoicissus digitata (NC_061712.1, https://www.ncbi.nlm.nih.gov/nuccore/NC_061712.1). The eight chloroplast genomes assembled by our research group in earlier stages of this project are not publicly available due to privacy reasons, they can be accessed through an online site (https://github.com/mrprinceq/Supplementary-data-1-8-eight-cpDNAs-).

Declarations

Ethics approval and consent to participate
All methods were performed in accordance with the relevant guidelines and regulations.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 8 July 2024 / Accepted: 19 August 2024
Published online: 02 September 2024

References

1. Quax TEF, Claassens NJ, Söll D, van der Oost J. Codon bias as a means to fine-tune gene expression. Mol Cell. 2015;59:149–61.
2. Tang D, Wei F, Cai Z, Wei Y, Khan A, Miao J, et al. Analysis of codon usage bias and evolution in the chloroplast genome of Mesona Chinensis Benth. Dev Genes Evol. 2021;231:1–9.
3. Angellotti MC, Bhuiyan SB, Chen G, Wan X-F. CodonO: codon usage bias analysis within and across genomes. Nucleic Acids Res. 2007;35 Web Server issue:W132–6.
4. Tuller T, Waldman YY, Kupiec M, Ruppin E. Translation efficiency is determined by both codon bias and folding energy. Proc Natl Acad Sci U S A. 2010;107:3645–50.
5. Sheng J, She X, Liu X, Wang J, Hu Z. Comparative analysis of codon usage patterns in chloroplast genomes of five Miscanthus species and related species. PeerJ. 2021;9:e12173.
6. Woodson JD. Control of chloroplast degradation and cell death in response to stress. Trends Biochem Sci. 2022;47:851–64.
7. Wang Z, Cai Q, Wang Y, Li M, Wang C, Wang Z, et al. Comparative analysis of Codon bias in the chloroplast genomes of Theaceae Species. Front Genet. 2022;13:824610.
8. Jiang H, Tian J, Yang J, Dong X, Zhong Z, Mwachala G, et al. Comparative and phylogenetic analyses of six Kenya Polystachya (Orchidaceae) species based on the complete chloroplast genome sequences. BMC Plant Biol. 2022;22:177.
9. Dobrogojski J, Adamiec M, Luciński R. The chloroplast genome: a review. Acta Physiol Plant. 2020;42:98.
10. Daniell H, Jin S, Zhu X-G, Gitzendanner MA, Soltis DE, Soltis PS. Green giant-a tiny chloroplast genome with mighty power to produce high-value proteins: history and phylogeny. Plant Biotechnol J. 2021;19:430–47.
11. Zhang P, Xu W, Lu X, Wang L. Analysis of codon usage bias of chloroplast genomes in Gynostemma species. Physiol Mol Biol Plants Int J Funct Plant Biol. 2021;27:2727–37.
12. Chakraborty S, Yengkhom S, Uddin A. Analysis of codon usage bias of chloroplast genes in Oryza species: Codon usage of chloroplast genes in Oryza species. Planta. 2020;252:67.
13. Nie X, Deng P, Feng K, Liu P, Du X, You FM, et al. Comparative analysis of codon usage patterns in chloroplast genomes of the Asteraceae family. Plant Mol Biol Rep. 2014;32:828–40.
14. Zhang L, Meng Y, Wang D, He G-H, Zhang J-M, Wen J, et al. Plastid genome data provide new insights into the dynamic evolution of the tribe Ampelopsideae (Vitaceae). BMC Genomics. 2024;25:247.
15. Wen J, Lu L, Nie Z, Liu X, Zhang N, Ickert-Bond S, et al. A new phylogenetic tribal classification of the grape family (Vitaceae). J Syst Evol. 2018;56:262–72.
16. Nho KJ, Chun JM, Kim D-S, Kim HK. Ampelopsis Japonica ethanol extract suppresses migration and invasion in human MDA–MB–231 breast cancer cells. Mol Med Rep. 2015;11:3722–8.
17. Oh Y, Lee H, Yang B, Kim S, Jeong H, Kim H. Anti-inflammatory effects of Ampelopsis Japonica Root on Contact Dermatitis in mice. Chin J Integr Med. 2022;28:719–24.
18. Zeng T, Song Y, Qi S, Zhang R, Xu L, Xiao P. A comprehensive review of vine tea: origin, research on Materia Medica, phytochemistry and pharmacology. J Ethnopharmacol. 2023;317:116788.
19. Carneiro RCV, Ye L, Baek N, Teixeira GHA, O'Keefe SF. Vine tea (*Ampelopsis grossedentata*): a review of chemical composition, functional properties, and potential food applications. J Funct Foods. 2021;76:104317.
20. Shi Y. Study on qualitative evaluation and teabag preparation technology of ethnomedicine vine tea. Master Thesis. Huazhong University of Science and Technology; 2022.
21. Gui C, Zou X, Yang X, Xu S, Liao S, Zhang X. Advances in studies on chemical compositions of and their biological activitie. J Hubei Univ Chin Med. 2023;25:118–21.
22. Zeng X, Liu T, Gong L, Xiao L, Wang W, Liu C et al. Pharmacognostical study of the Ethnomedicinal Plant Ampelopsis grossedentata var. Microphylla. J Chin Med Mater. 2023;:2172–7.
23. Luo Y. Comparative analyses and phylogenetic relationships of complete chloroplast genomes from the Ampelopsideae (Vitaceae). Master Thesis. Hubei Minzu University; 2024.
24. Yan F, Jiang R, Wang L, Luo Z. Codon Preference Analysis of Chloroplast Genome in Dendrobium Cariniferum. Guizhou Sci Technol. 2024;52:29–36.
25. Edelman GM, Gally JA. Degeneracy and complexity in biological systems. Proc Natl Acad Sci U S A. 2001;98:13763–8.
26. Wang L, Xing H, Yuan Y, Wang X, Saeed M, Tao J, et al. Genome-wide analysis of codon usage bias in four sequenced cotton species. PLoS ONE. 2018;13:e0194372.
27. Mao L, Huang Q, Long L, Tan X, Xie H, Tang Y, et al. Comparative Analysis of Codon Usage Bias in Chloroplast Genomes of Seven Nymphaea Species. J Northwest Univ. 2022;37:98–107.
28. Chen J, Wang F, Zhao Z, Li M, Liu Z, Peng D. Complete chloroplast genomes and comparative analyses of three paraphalaenopsis (Aeridinae, Orchidaceae) Species. Int J Mol Sci. 2023;24:11167.
29. He J, Xu K, Du Y, Jiang H, Niu Q, Wang Z et al. Progress in research on mitochondrial genome of honey bees and their polymorphisms. J Environ Entomol. 2024;:1–18.

30. Li Q, Luo Y, Sha A, Xiao W, Xiong Z, Chen X, et al. Analysis of synonymous codon usage patterns in mitochondrial genomes of nine Amanita species. Front Microbiol. 2023;14:1134228.
31. Yan C, Xu Q, Li Z, Ran Z. Characterization of Complete Chloroplast Genome and Phylogenetic Analysis of Camellia lipingensis. Mol Plant Breed. 2024;:1–22.
32. Zeng Y, Shen L, Chen S, Qu S, Hou N. Codon usage profiling of Chloroplast Genome in Juglandaceae. Forests. 2023;14:378.
33. Luo Y, Wang R, Zhao R, Lu X, Yin G, Deng Z. Analysis of synonymous codon usage bias in the chloroplast genome of Davidia involucrata. J BEIJING Univ. 2024;46:8–16.
34. Wang Z, Xu B, Li B, Zhou Q, Wang G, Jiang X, et al. Comparative analysis of codon usage patterns in chloroplast genomes of six Euphorbiaceae species. PeerJ. 2020;8:e8251.
35. Zhou Z, Dang Y, Zhou M, Li L, Yu C-H, Fu J, et al. Codon usage is an important determinant of gene expression levels largely through its effects on transcription. Proc Natl Acad Sci U S A. 2016;113:E6117–25.
36. Zhou M, Tong C-F, Shi J-S. A preliminary analysis of synonymous codon usage in poplar species. Zhi Wu Sheng Li Yu Fen Zi Sheng Wu Xue Xue Bao. 2007;33:285–93.
37. Sahoo S, Das SS, Rakshit R. Codon usage pattern and predicted gene expression in Arabidopsis thaliana. Gene X. 2019;2:100012.
38. Hershberg R, Petrov DA. General rules for optimal codon choice. PLoS Genet. 2009;5:e1000556.
39. Boël G, Letso R, Neely H, Price WN, Wong K-H, Su M, et al. Codon influence on protein expression in E. Coli correlates with mRNA levels. Nature. 2016;529:358–63.
40. Taunt HN, Stoffels L, Purton S. Green biologics: the algal chloroplast as a platform for making biopharmaceuticals. Bioengineered. 2018;9:48–54.
41. Itaya H, Oshita K, Arakawa K, Tomita M. GEMBASSY: an EMBOSS associated software package for comprehensive genome analyses. Source Code Biol Med. 2013;8:17.
42. Singh NK, Tyagi A. A detailed analysis of codon usage patterns and influencing factors in Zika virus. Arch Virol. 2017;162:1963–73.
43. Chi X, Zhang F, Dong Q, Chen S. Insights into Comparative Genomics, Codon usage Bias, and phylogenetic relationship of species from Biebersteiniaceae and Nitrariaceae based on complete chloroplast genomes. Plants Basel Switz. 2020;9:1605.
44. Wu Y, Li Z, Zhao D, Tao J. Comparative analysis of flower-meristem-identity gene *APETALA2* (*AP2*) codon in different plant species. J Integr Agric. 2018;17:867–77.
45. Wang Y, Jiang D, Guo K, Zhao L, Meng F, Xiao J, et al. Comparative analysis of codon usage patterns in chloroplast genomes of ten Epimedium species. BMC Genomic Data. 2023;24:3.
46. Sueoka N. Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci. 1988;85:2653–7.
47. Gao Y, Lu Y, Song Y, Jing L. Analysis of codon usage bias of WRKY transcription factors in Helianthus annuus. BMC Genomic Data. 2022;23:46.
48. Wright F. The effective number of codons used in a gene. Gene. 1990;87:23–9.
49. Chen S, Zhang H, Wang X, Zhang Y, Ruan G, Ma J. Analysis of Codon usage Bias in the chloroplast genome of Helianthus annuus J-01. IOP Conf Ser Earth Environ Sci. 2021;792:012009.
50. Sueoka N. 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. 1999;238:53–8.
51. Zhang Y, Nie X, Jia X, Zhao C, Biradar SS, Wang L, et al. Analysis of codon usage patterns of the chloroplast genomes in the Poaceae family. Aust J Bot. 2012;60:461.
52. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform. 2019;20:1160–6.
53. Nguyen L-T, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32:268–74.
54. Yang M, Liu J, Yang W, Li Z, Hai Y, Duan B, et al. Analysis of codon usage patterns in 48 Aconitum species. BMC Genomics. 2023;24:703.

Publisher's note