forests

MDPI

Article

Comparative Analysis of Codon Usage Patterns in Chloroplast Genomes of Cherries

Yan-Feng Song [1,2], Qing-Hua Yang [1,2], Xian-Gui Yi [1,2], Zhao-Qing Zhu [3], Xian-Rong Wang [1,2] and Meng Li [1,2,4,*]

[1] Co-Innovation Center for Sustainable Forestry in Southern China, College of Biology and the Environment, Nanjing Forestry University, Nanjing 210037, China
[2] Cerasus Research Center, College of Biology and the Environment, Nanjing Forestry University, Nanjing 210037, China
[3] Shanghai Jishi Landscape Co., Ltd., Shanghai 200080, China
[4] Key Laboratory of National Forestry and Grassland Administration on Subtropical Forest Biodiversity Conservation, College of Biology and the Environment, Nanjing Forestry University, Nanjing 210037, China
* Correspondence: limeng@njfu.edu.cn

Abstract: Synonymous codon usage bias (SCUB) analysis is an effective method to explore species specificity, evolutionary relationships and mRNA translation, as well as to discover novel genes, which are important for understanding gene function and molecular phylogeny. Cherries (*Prunus* subg. *Cerasus*) are flowering plant germplasm resources for edible and ornamental purposes. In this study, we analyzed the codon usage patterns of the 36 chloroplast genomes to provide a scientific basis for elucidating the evolution of subg. *Cerasus*. The results showed that the average GC content was 0.377, the average GC3 was 0.298, and the average ENC value was 49.69. Neutral-plot analysis, ENC-plot analysis, and PR2-plot analysis all indicated that natural selection was the main factor of codon usage bias in subg. *Cerasus*, whereas correlation analysis showed that gene expression level and GC1 also affect the codon usage pattern. The codon usage pattern was consistent across 36 species, and 30 high-frequency codons were identified, with preference for A/T endings; there were 23 optimal codons, and only GAU was identified in all individuals; structural differences existed between the clustering tree based on RSCU values and the phylogenetic tree based on CDS, elucidating the importance of locus mutations and no-preference codons in phylogenetic reconstruction. This study describes for the first time the SCUB pattern and characterization of subg. *Cerasus* chloroplast genomes and provides a new insight to explore the phylogeny of this subgenus.

Keywords: *Prunus*; *Cerasus*; plastid genome; codon usage; phylogeny

Citation: Song, Y.-F.; Yang, Q.-H.; Yi, X.-G.; Zhu, Z.-Q.; Wang, X.-R.; Li, M. Comparative Analysis of Codon Usage Patterns in Chloroplast Genomes of Cherries. *Forests* **2022**, *13*, 1891. https://doi.org/10.3390/f13111891

Academic Editor: Tadeusz Malewski

Received: 22 September 2022
Accepted: 5 November 2022
Published: 10 November 2022

1. Introduction

The triplet codons perform a fundamental role in protein translation, and the transfer of genetic information from mRNA to protein is a basic link in maintaining the viability of the organism. All amino acids are encoded by two to six codons, except for methionine and tryptophan, which are encoded by only one codon [1,2]. Codons encoding the same amino acid are considered as synonymous codons. In the matter of the process of protein translation, different species tended to use one or more specific synonymous codons called synonymous codon usage bias (SCUB) [3,4]. According to previous studies, protein translation efficiency, tRNA abundance, gene drift, and expression level have all been associated with SCUB, but natural selection and genes mutation are the most important factors [5–9].

Chloroplast is the main organelle used to execute photosynthesis activities and metabolic reactions in green plants. Compared with the nuclear genome, the chloroplast genome in angiosperms is a circular DNA molecule which has a highly conserved genomic structure with a small size, single-parental inheritance, and low nucleotide substitution rate; hence, it is widely used as molecular evidence for phylogenetic analysis and

species identification [10–14]. With the rapid development of high-throughput sequencing technology, massive plants taxon chloroplast genomes have been sequenced, offering the research of chloroplast genome codon with a database.

With the popularity of next-generation genome sequencing technologies, the study of plant genome codons is increasing. However, most studies on plant codon preferences have been conducted only in single species, and comparison of interspecific codon usage patterns is lacking. Cherry is a generic term for *Prunus* subg. *Cerasus* species, which contains approximately 50 species that mainly occupy temperate and subtropical regions of the northern hemisphere [15–19]. Cherries offer a variety of edible drupes and ornamental plants with economic value to human society and thus have great potential for development and application. However, frequent interspecific hybridization and sympatric speciation have yielded little knowledge of the phylogeny of subg. *Cerasus* or the evolutionary forces driving the evolutionary process. At present, numerous studies on the synonymous codon preference of plant chloroplast genomes have been reported, such as Poaceae [20], Magnoliaceae [21], Asteraceae [22], Euphorbiaceae [13], Fragaria [23], and others. In this study, we combined 24 published chloroplast genomes of subg. *Cerasus* and 12 self-assembled plastomes to systematically analyze the synonymous codon usage patterns of subg. *Cerasus* species and reveal the phylogenetic relationships of this taxon, aiming to provide new insights into the evolution of the cherry chloroplast genome as well as scientific information and stable data for the application and conservation of germplasm resources basis.

2. Materials and Methods

2.1. Sampling, DNA Isolation and Sequencing

In order to address the issue of unrepresentative samples in previous studies [24,25], we newly sampled and sequenced the plastome of 12 subg. *Cerasus* species in this study, namely *Prunus clarofolia* (31°20′52.52″ N, 109°58′46.35″ E), *P. conadenia* (28°39′36.84″ N, 97°27′54.05″ E), *P. discoidea* (30°4′14.08″ N, 118°5′25.54″ E), *P. jamasakura* (34°45′53.7″ N, 135°42′10.5″ E), *P. mahaleb* (42°18′23.58″ N, 71°7′19.38″ W), *P. mugus* (27°35′52.81″ N, 98°39′9.62″ E), *P. polytricha* (31°19′29.22″ N, 109°59′2.76″ E), *P. sargentii* (42°9′17.21″ N, 140°9′57.19″ E), *P. schneideriana* (28°18′31.76″ N, 119°18′6.41″ E), *P. serrula* (25°46′30.23″ N, 102°20′30.58″ E), *P. setulose* (34°29′33.52″ N, 103°40′6.34″ E) and *P. yunnanenesis* (23°40′8.88″ N, 106°12′46.74″ E). These fresh mature leaves collected from robust individuals of the wild subg. *Cerasus* species were packed into labeled tea bags and immediately placed in silica gel for drying and storage.

Total Genomic DNA was extracted from each of subg. *Cerasus* plants using a DNeasy Plant Mini Kit (Qiagen Co., Hilden, Germany) following the manufacturer's protocol. The extracted DNA was quantified in NanoDrop ND1000 (Thermo Fisher Scientific, Waltham, MA, USA; quality cutoff, OD 260/280 ratio between 1.7–1.9) and visualized in a 1% agarose-gel electrophoresis for the quality check. Illumina paired-end (PE) libraries (read length: 2 × 125 bp) with insert sizes of 270 to 700 bp for each of subg. *Cerasus* species were constructed and sequenced on MiSeq platform (Illumina Inc., San Diego, CA, USA) by Nanjing Genepioneer Biotechnologies Inc. (Nanjing, China). We removed poor-quality reads (PHRED score of <20) using the quality trim function implemented in CLC Assembly Cell package v. 4.2.1 (CLC Inc., Aarhus, Denmark).

The average clean PE reads for each sample was 3.54 Gb (Phred score > 20), and we subsequently assembled de novo these clean reads using the GetOrganelle pipeline [26]. The assembled chloroplast genomes were annotated by the web application GeSeq (https://chlorobox.mpimp-golm.mpg.de/geseq.html (accessed on 29 April 2022) [27]. Another 24 subg. *Cerasus* plastomes were downloaded from the National Center for Biotechnology Information (NCBI). All samples are listed in Table S1. The coding sequences (CDS) of each species were extracted by Phylosuite v1.2.2 [28].

We filtered coding sequences less than 300 bp and removed duplicate genes and stop codons, and a total of 58 shared CDSs in 36 subg. *Cerasus* species were enumerated for subsequent analysis.

2.2. Codon Composition

We analyzed RSCU (relative synonymous codon usage), ENC (effective number of codons) and CAI (codon adaptation index) of selected protein-coding genes using CodonW v1.4.2 (http://codonw.sourceforge.net/ (accessed on 3 May 2022) and calculated the GC content of the first base of codon (GC1), the second base of codon (GC2), the third base of codon (GC3), and total GC content (GC). The RSCU were calculated as Equation (1):

$$RSCU = \frac{x_{ij}}{\sum_j^{n_i} x_{ij}} n_i \quad (1)$$

where x_{ij} represents the frequency of codon j encoding the i-th amino acid and n_i is the number of synonymous codons encoding the i-th amino acid. If the RSCU value of a codon is equal to 1, it reflects no codon usage bias, instead, the RSCU value reflects the frequency of usage deviation from other codons.

2.3. Neutrality Plot

In the neutral graph, the average values of GC1 and GC2 of each gene were regarded as GC12, which were ordinate; GC3 values were seen as abscissa. Finally, we line-fit the scatter points in the diagram. Often, when the regression coefficient is closer to 1, the SCUB is more affected by the base mutation; the number closer to 0 is greater affected by natural selection.

2.4. Analysis of ENC-Plot

ENC values are used to analyze the extent to which codon usage deviates from random selection, which depicts the extent to which synonymous codons are used unevenly in a specific gene or genome in a given species, ranging from 20 to 61. If the value is greater than 35, it indicates that the use of codons has high bias and vice versa. We plotted scatters with ENC content as ordinate and GC3 values as horizontal coordinates and draw standard curve. The expected values of *ENC* were calculated according to Equation (2):

$$ENC = 2 + GC3s + \frac{29}{GC3s^2 + (1 - GC3s^2)} \quad (2)$$

When mutational pressure has a significant effect on codon usage patterns, ENC values lie on or near the expected curve. On the contrary, when influenced by natural selection and other factors, ENC values are well below the expected curve.

2.5. PR2-Plot

Previous studies have witnessed that the third base of the codon is a critical factor affecting its preference; we chose the third base of the codon to A3/(A3 + T3) as the ordinate, and G3/(G3 + C3) for the horizontal coordinates to draw a PR2-plot diagram to explore whether its preference is related to natural selection or other factors [29].

2.6. Correlation Analysis

Correspondence analysis comparing the usage patterns of 59 codons (excluding codons encoding Met, Trp and three stop codons) led to a series of orthogonal axes that can be used to demonstrate codon usage variation in the chloroplast genome of cherries. Based on SCUB of genes in a multidimensional space of 59 axes, it is possible to derive the distribution of genes, as well as the largest fraction of gene variation, and in this regard to analyze the main factors of codon variation. The codon usage variations in 36 subg. *Cerasus* species were investigated with correspondence analysis based on RSCU using CodonW v1.4.2 [30].

The correlation between the first four axes and each codon usage parameters was analyzed by SPSS v 25.0.

2.7. Optimal Codons

We sorted the 58 protein coding sequences that were filtered from highest to lowest ENC values, then selected the first 10% and the last 10% genes to establish a high–low bias library and calculated the ΔRSCU value (RSCU high − RSCU low). If the RSCU > 1 and ΔRSCU > 0.08, then this codon was defined as high expression superior codon for this species [31].

2.8. Cluster and Phylogenetic Analysis

In order to explore the relationship between the SCUB of plastomes in subg. *Cerasus* plants and its phylogenetic relationship, we conducted a cluster analysis of 36 subg. *Cerasus* in R v3.6.3 using RSCU values as variables, calculated as "Euclidean distance," and plotted the cluster spectrum [32]. We also reconstructed a phylogenetic tree based on 58 CDS sequences using a maximum likelihood method with 1000 bootstrap replications [33]. The tested outgroups were *Ulmus chenmoui* and *Ziziphus jujuba*, both downloaded from NCBI (Genbank accession number: MG581403 and KU351660).

3. Results

3.1. Base Composition Characteristics

As shown in Table 1, the average number of codons in 36 species was 26,165, of which *P. conadenia* had the lowest number (25,205) and *P. pseudocerasus* had the highest number (26,575). The average content of GC1, GC2, GC3 and GC content of CDS sequences were 0.454, 0.377, 0.298 and 0.377 respectively. The results indicated that the three codon bases were biased towards A (adenine) and T (thymine). The GC contents of all tested chloroplast genomes were consistent with GC1 > GC2 > GC3 (Figure 1). The average ENC value was 49.69, the minimum ENC value of *P. fruticosa* was 49.61, and the highest value of *P. discoidea* was 49.76 (Figure 2). All ENC values of this taxon were greater than 49, reflecting weak codon bias overall. The CAI value of all species was 0.166, except *P. conadenia* which was 0.167.

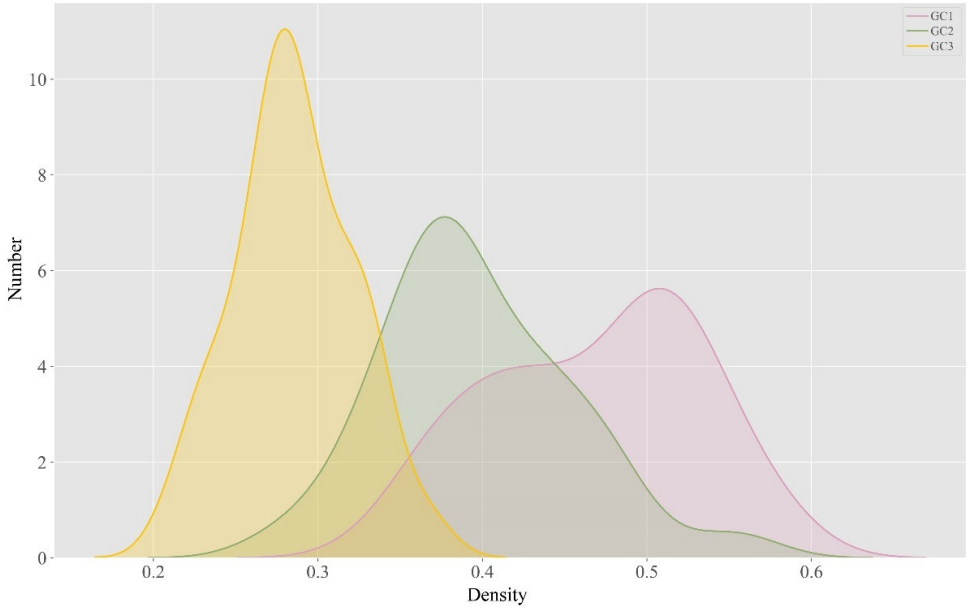

Figure 1. The kernel density curve of GC1, GC2 and GC3. The x-axis represents the density of GC1, GC2 and GC3 contents in all test CDS, and the y-axis indicates the relative quantity.

Table 1. Codon features of 36 subg. *Cerasus* chloroplast genomes.

Species	Codon No.	CG1	CG2	CG3	GC	ENC	CAI
Prunus apetala	26158	0.4529	0.3765	0.2982	0.3760	49.71	0.166
Prunus avium	26162	0.4541	0.3770	0.2978	0.3770	49.66	0.166
Prunus campanulata	26157	0.4542	0.3770	0.2984	0.3770	49.69	0.166
Prunus cerasoides	26172	0.4542	0.3767	0.2979	0.3770	49.68	0.166
Prunus clarofolia	26209	0.4545	0.3768	0.2981	0.3770	49.69	0.166
Prunus conadenia	25205	0.4560	0.3762	0.2974	0.3770	49.67	0.167
Prunus conradinae	26490	0.4535	0.3762	0.2988	0.3770	49.73	0.166
Prunus dielsiana	26151	0.4543	0.3769	0.2984	0.3770	49.70	0.166
Prunus discoidea	26525	0.4533	0.3760	0.2991	0.3770	49.76	0.166
Prunus emarginata	26163	0.4542	0.3769	0.2976	0.3770	49.66	0.166
Prunus fengyangshanica	26163	0.4542	0.3767	0.2984	0.3770	49.71	0.166
Prunus fruticosa	26166	0.4537	0.3766	0.2971	0.3760	49.61	0.166
Prunus itosakura	26152	0.4546	0.3770	0.2982	0.3770	49.69	0.166
Prunus jamasakura	26160	0.4541	0.3769	0.2982	0.3770	49.68	0.166
Prunus jingningensis	26165	0.4542	0.3770	0.2983	0.3770	49.69	0.166
Prunus kumanoensis	26158	0.4543	0.3770	0.2982	0.3770	49.68	0.166
Prunus leveilleana	26158	0.4543	0.3770	0.2983	0.3770	49.69	0.166
Prunus mahaleb	26218	0.4547	0.3771	0.2980	0.3770	49.67	0.166
Prunus matuurae	26156	0.4540	0.3769	0.2986	0.3770	49.72	0.166
Prunus maximowiczii	26158	0.4543	0.3768	0.2984	0.3770	49.71	0.166
Prunus mugus	25990	0.4552	0.3775	0.2977	0.3770	49.67	0.166
Prunus pensylvanica	26162	0.4544	0.3769	0.2980	0.3770	49.68	0.166
Prunus polytricha	26220	0.4543	0.3769	0.2984	0.3770	49.70	0.166
Prunus pseudocerasus	26575	0.4535	0.3761	0.2988	0.3770	49.73	0.166
Prunus rufa	26171	0.4541	0.3769	0.2979	0.3770	49.67	0.166
Prunus sargentii	26153	0.4543	0.3770	0.2981	0.3770	49.67	0.166
Prunus schneideriana	26224	0.4543	0.3769	0.2985	0.3770	49.70	0.166
Prunus serrula	26217	0.4544	0.3769	0.2983	0.3770	49.69	0.166
Prunus setulosa	26227	0.4539	0.3769	0.2980	0.3770	49.67	0.166
Prunus speciosa	26164	0.4540	0.3769	0.2983	0.3770	49.69	0.166
Prunus spontanea	26158	0.4543	0.3770	0.2984	0.3770	49.70	0.166
Prunus subhirtella	26152	0.4546	0.3770	0.2982	0.3770	49.69	0.166
Prunus takesimensis	26158	0.4543	0.3770	0.2983	0.3770	49.69	0.166
Prunus verecunda	26158	0.4544	0.3770	0.2983	0.3770	49.69	0.166
Prunus yedoensis	26152	0.4546	0.3770	0.2982	0.3770	49.69	0.166
Prunus yunnanensis	26003	0.4551	0.3774	0.2977	0.3770	49.66	0.166

3.2. Neutrality Plot Analysis

In order to determine the main factors influencing the SCUB of chloroplast genome, we carried out a neutral plot analysis of 36 subg. *Cerasus* species. As shown in Figure 3, GC12 and GC3 were scattered distribution. GC12 was 0.2852–0.5432, and GC3 was 0.2074–0.3661. The regression coefficient was between 0.1331 and 0.2455, where *P. conradinae* was 0.1331, which was most influenced by natural selection; *P. conadenia* was 0.2455, which was least affected by natural selection. These results suggested that the SCUB of subg. *Cerasus* chloroplast genomes were mainly influenced by natural selection.

Neutrality plot analysis reflects only the factors that influence codon usage patterns; consequently, further analysis of the degree of influence of mutational pressure and natural selection is necessary.

3.3. ENC-Plot Analysis, PR2-Plot Analysis and Correlation Analysis

As shown in Figure 4a, the genes of 36 subg. *Cerasus* species were scattered on both sides of the expected curve, with most genes not near or below the expected curve and a few genes with ENC values less than 35. These results also demonstrated that natural selection is the dominant factor of SCUB of chloroplast genome in subg. *Cerasus* plants.

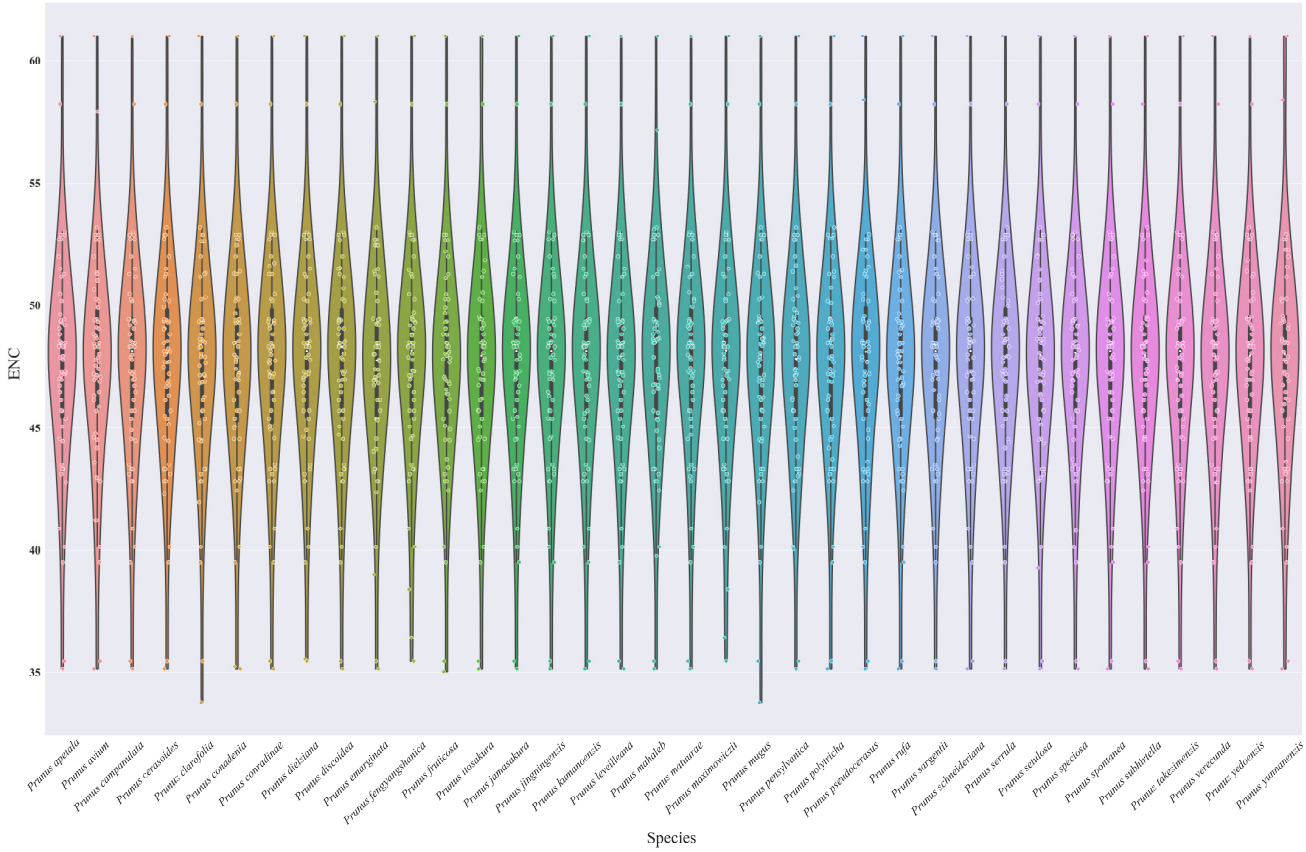

Figure 2. The ENC values of chloroplast genes in subg. *Cerasus* species. The x-axis represents 36 sub. *Cerasus* accessions, and the y-axis indicates the ENC value for all CDS in each species.

We performed a parity check analysis on the relationship between the third base of the codons of all the tested genes of 36 subg. *Cerasus* plants (Figure 4b). The diagram displayed that tested genes were unevenly distributed in four regions, with a large number of genes located in the bottom right, which meant that T was used more frequently than A, and G (guanine) was greater than C (cytosine). Overall, the frequency of A/T and G/C use in the chloroplast genomes of subg. *Cerasus* was asymmetrical and was influenced not only by mutational pressure as well as by natural selection and other factors. This result further validated the effect of natural selection on SCUB in the cherries.

We analyzed the correspondence patterns of 36 subg. *Cerasus* chloroplast genes based on RSCU values. The Axis 1 was the major factor in causing the variation, responsible for about 10% of total variation in all tested subg. *Cerasus* species, with each subsequent axis explaining a decreasing amount of the variation. Based on Pearson correlation analysis, the Axis 1 was significantly correlated with GC1, GC3, GC12, GC, CAI and ENC ($p < 0.01$) (see Table 2), which means that the base composition greatly affected the codon usage bias. GC1 and CAI value were likely to affect the patterns of codon usage, and significantly correlated with COA axes (except Axis 3). This suggested that the first base of codon and the gene expression level were likely involved in some subg. *Cerasus* chloroplast genomes. Furthermore, there were no obvious characteristics of the distribution of subg. *Cerasus* chloroplast genes (Figure 4c).

3.4. High-Frequency Codon and Optimal Codon

The RSCU value of 36 subg. *Cerasus* genome sequences was calculated using CodonW, and high-frequency codons (RSCU > 1) were counted (Supplementary Table S2 and Figure 5). In 36 tested plants, except for the UUG encoded L (leucine) ending in G, the other codons ended in A (13) or T (16), indicating that codon ending in A/T were used more frequently. At amino acid levels, UUA, which also encodes L, showed strong preference in

all subg. *Cerasus* species. We also sorted the ENC values of all tested chloroplast genes in 36 species, picked out five high-value and five low-value genes to build a library, and then selected the codons that met the conditions as the optimal codons (Supplementary Table S3 and Figure 6).

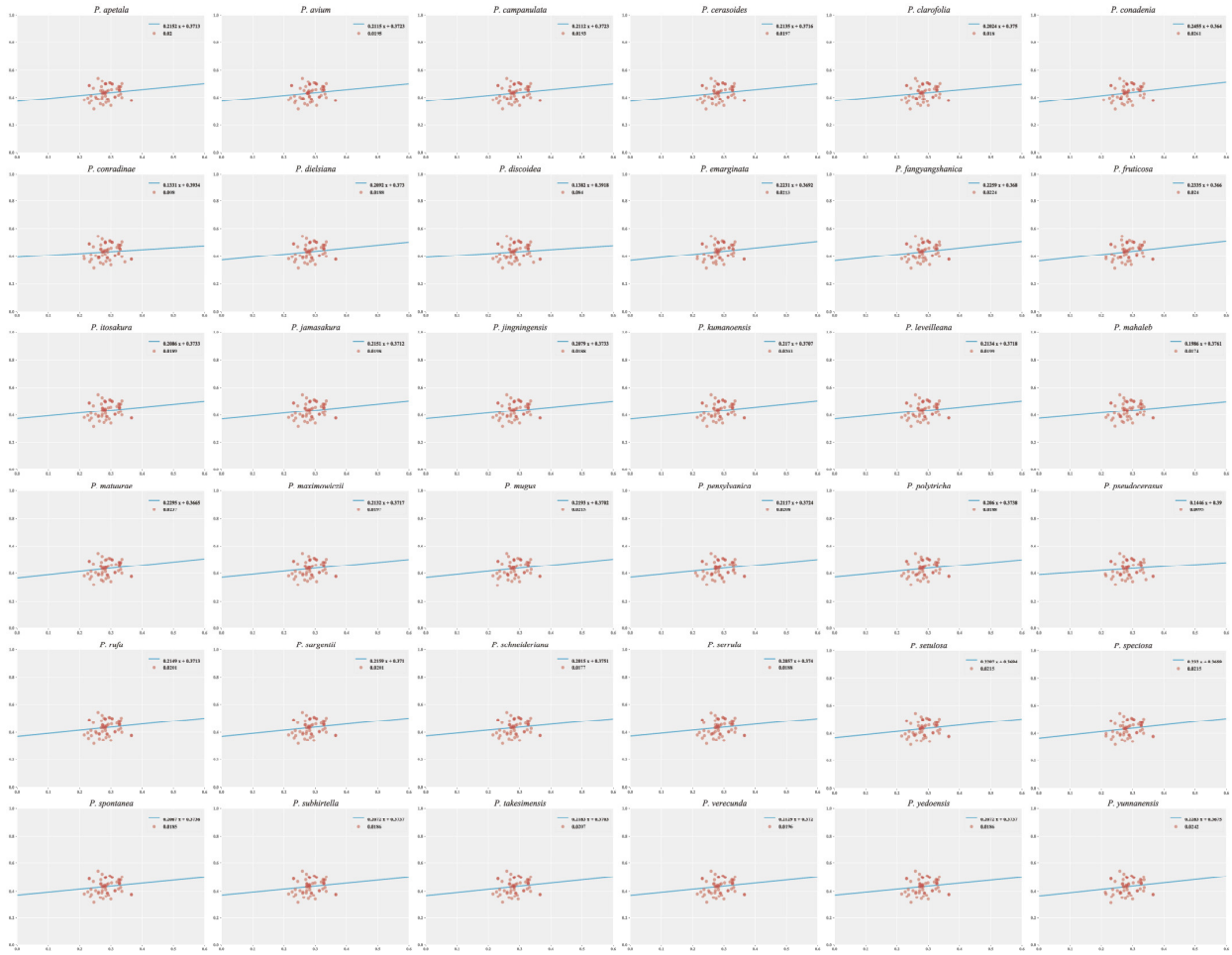

Figure 3. Neutrality plot analysis of 36 subg. *Cerasus* species.

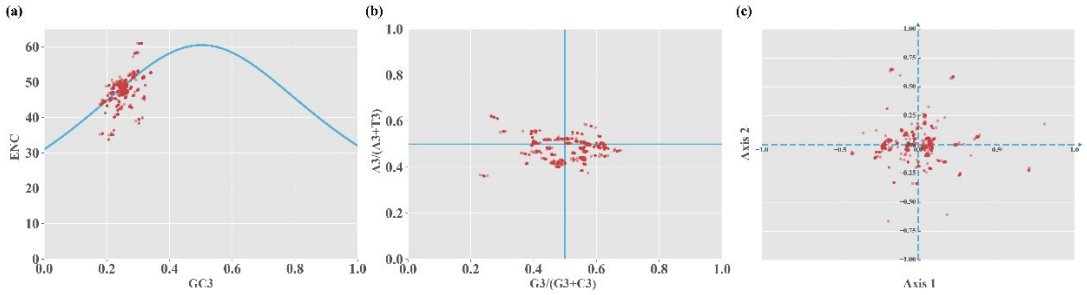

Figure 4. Related plotting analysis of codon usage frequency, with red dots representing protein-coding genes of 36 subg. *Cerasus* species. (**a**) ENC-plot analysis based on the concatenated coding sequences, the x-axis represents the GC3 value of CDS and the y-axis represents its ENC value; (**b**) PR2-plot analysis, the x-axis denotes the value of G3/(G3 + C3) and the y-axis denotes the value of A3/(A3 + T3); (**c**) correspondence analysis of the RSCU values of subg. *Cerasus* species, each point on the plot corresponds to the coordinates on the first and second-principal axes produced by the COA.

Table 2. The correlation coefficients between the axes and codon usage indices of chloroplast genes in 36 subg. *Cerasus* species.

	GC1	GC2	GC3	GC12	GC	CAI	ENC
Axis 1	−0.141 **	0.035	−0.103 **	−0.066 **	−0.091 **	−0.130 **	−0.071 **
Axis 2	−0.074 **	0.026	−0.030	−0.030	−0.037	0.147 **	−0.055 *
Axis 3	−0.023	0.042	−0.023	0.009	0.002	−0.050 *	−0.107 **
Axis 4	−0.050 *	−0.098 **	0.094 **	−0.084 **	−0.049 *	0.077 **	0.034

Notes. * $p < 0.05$. ** $p < 0.01$.

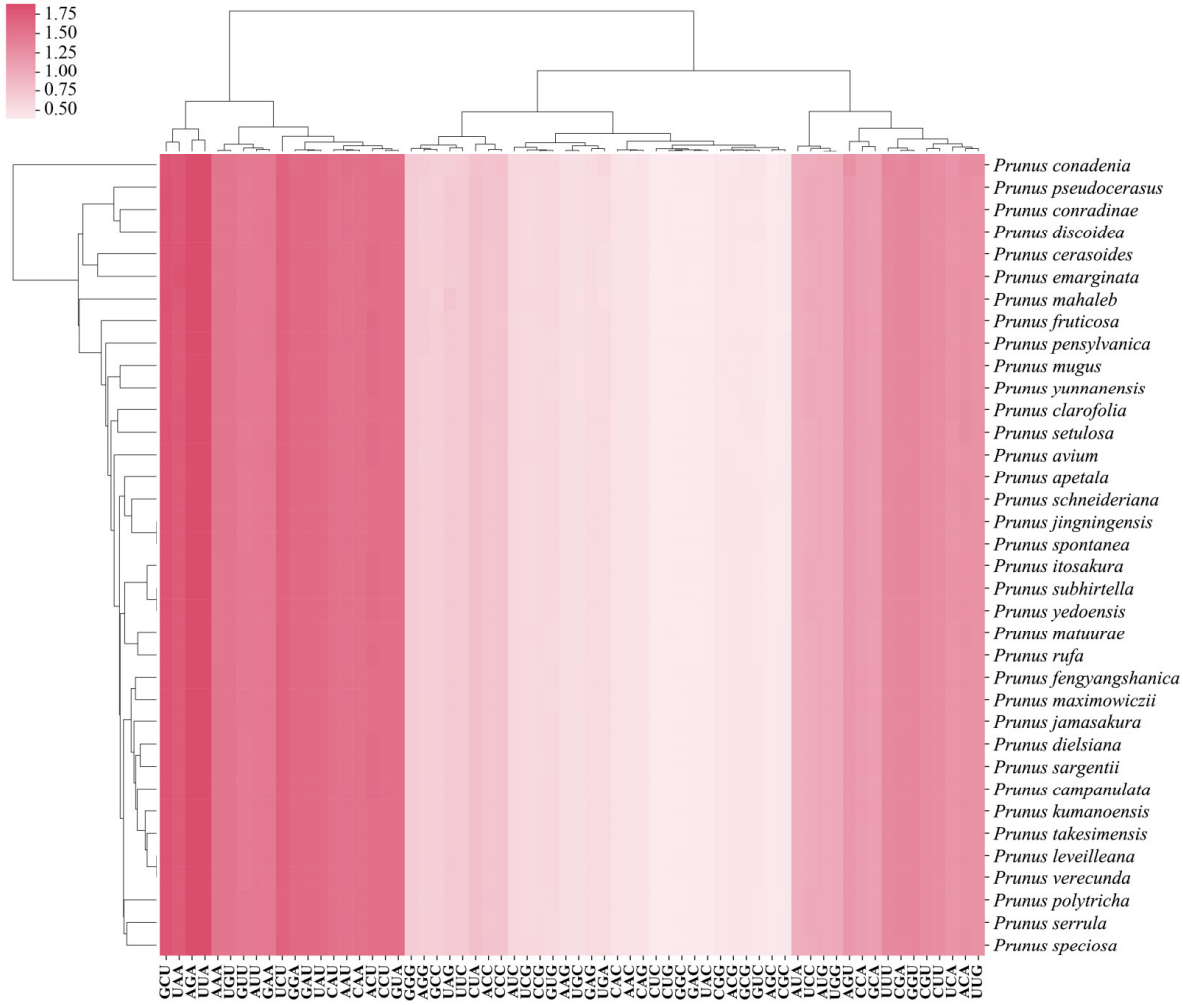

Figure 5. The codon usage of 36 subg. *Cerasus* based on RSCU values of chloroplast genes. The x-axis indicates the triplet codons, and the y-axis represents the 36 subg. *Cerasus* accessions.

A total of 23 optimal codons encoded 12 amino acids were counted, among which 17 ended in A/T, accounting for 73.9%. The results manifested that there were significant differences in the optimal codons of 36 subg. *Cerasus* chloroplast genomes. *P. emarginata* had the most optimal codons, which was 11, while *P. itosakura* had the least number, which was 6. Of 12 species with 7 optimal codons, *P. cerasoides*, *P. kumanoensis*, *P. leveilleana*, *P. rufa*, *P. subhirtella* and *P. yedoensis* harbored the same composition of optimal codons, which were: ACC, AGU, CCU, CGU, GAU, GCC, GCU, UCC and UCU. Among 14 species with 8 optimal codons, ACC, AGA, AGU, CCU, CGU, GAU, GCC and UCU were the optimal codons of *P. clarofolia* and *P. mugus*. Secondly, *P. fengyangshanica*, *P. jingningensis*, *P. matuurae*, *P. maximowiczii*, *P. sargentii*, *P. speciosa*, *P. spontanea*, *P. takesimensis* and *P. verecunda* had the same optimal codon composition, but the difference from the former was that AAU

replaced AGA. Moreover, only one GAU encoding aspartic acid was the common optimal codon of all subg. *Cerasus* species.

Figure 6. Optimal codon analysis of 36 subg. *Cerasus* species. The x-axis indicates the optimal codon and its corresponding encoded protein, and the y-axis represents the 36 subg. *Cerasus* accessions. Dark block positions indicate optimal codons in each species.

3.5. Cluster and Phylogenetic Analysis

Based on the RSCU values of 36 subg. *Cerasus* plastome genomes, we constructed a hierarchical diagram (Supplementary Table S2 and Figure 7a). The result of cluster analysis revealed that 36 subg. *Cerasus* species were divided into 4 clusters, of which only one species in Cluster IV was *P. conadenia*, and the other three clusters contained 10, 22, and 3 species, respectively. Intriguingly, in contrast to the arborescence structure of the cluster analysis, the CDS-based phylogeny split into five clades exhibiting a completely different topology (Figure 7b). *P. mahaleb* placed basal in subg. *Cerasus* and formed a separate branch that was Clade I; Clade II contained 2 species, namely *P. rufa* and *P. cerasoides*; Clade III and Clade IV included 5 and 7 species, respectively; Clade V gathered the largest number of species of wild cherries, with 21.

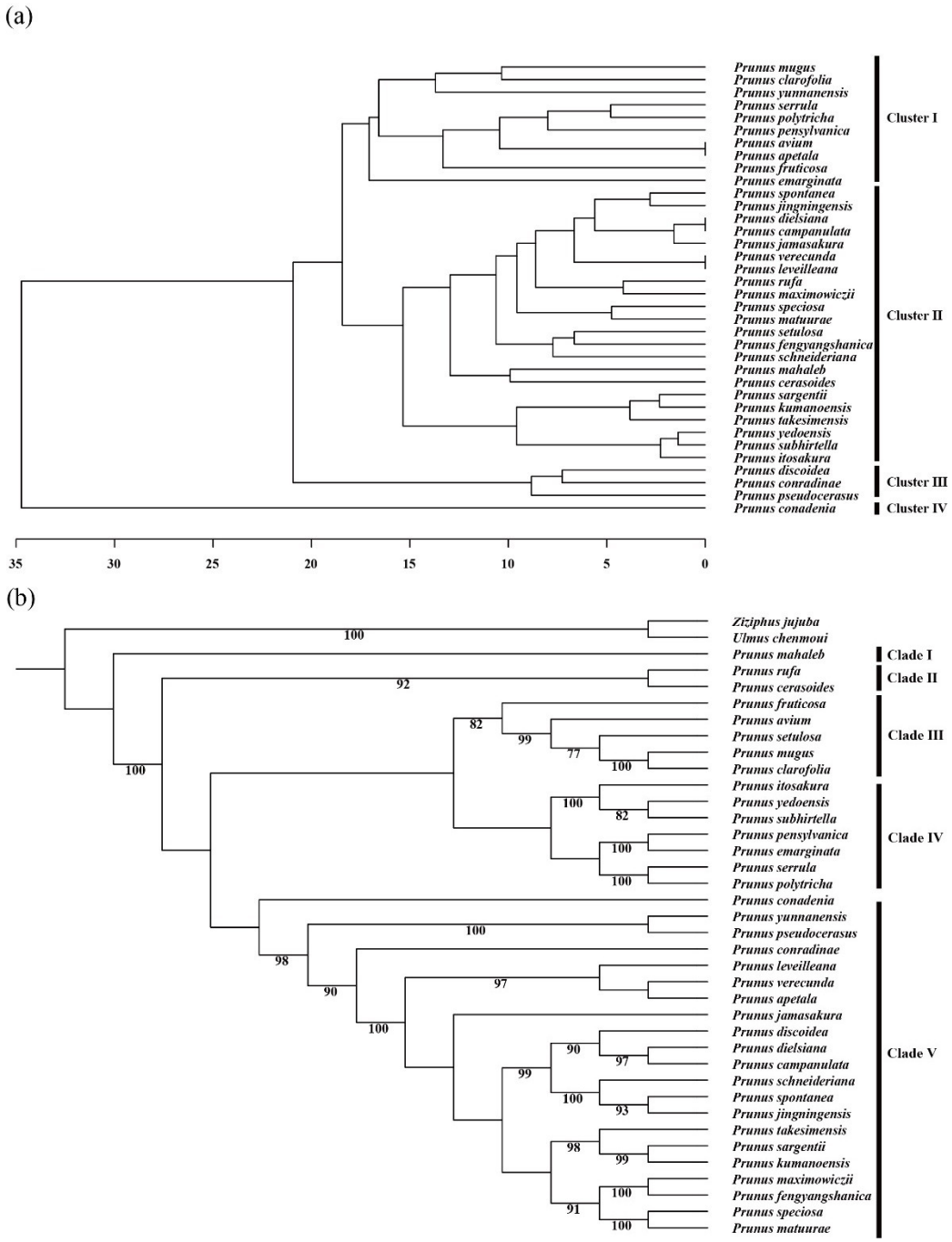

Figure 7. (**a**) Cluster of 36 subg. *Cerasus* chloroplast genomes based on RSCU values; (**b**) CDS-based reconstruction of a phylogenetic tree of subg. *Cerasus*.

4. Discussion

In general, the codon usage bias not only reflects the origin, evolution, and mutation patterns of species or genes, but also has conspicuous implications for gene function and expression [8,13,34]. The substitution in the third base of codon may not instantly alter the corresponding encoded amino acids, but they directly reflect the patterns of codon usage [35,36]. Our analysis results were consistent with most angiosperms codon bias: the GC content of subg. *Cerasus* chloroplast genome was 0.377 and the preferred ending with base A/T; all tested species appeared to have a weak codon preference (ENC value > 49) (Table 1 and Figure 2) [13,22,37]. In case codon preference is affected by natural selection, GC3 value tend to be distributed in a small range in the diagram, and GC12 have no

significant correlation with GC3. We observed a regression line with a lower slope in the 36 species, and also found the regression coefficients were close to 0 (Figure 3), which suggested that natural selection played an important role in codon usage bias in chloroplast genomes of subg. *Cerasus* plants. ENC-plot and PR2-plot analysis both proved that natural selection was the main factor influencing codon preference (Figure 4a,b). Apart from mutations and natural selection, we demonstrated that gene expression levels and the first base of codon also affected codon preferences through COA based on RSCU values (Table 2 and Figure 4c) [38]. 36 subg. *Cerasus* genomes encompassed 30 identical high-frequency codons, and the third base was generally biased towards A/T (Table S2 and Figure 5). On the contrary, the number of interspecies optimal codons ranged 6 from 11 and the only shared codon was GAU (Figure 6). Therefore, the reason for the difference between optimal codon and synonymous codon was that the former was affected by ENC and expression level of different genes.

The results of gene annotation in this study indicated that the least number of codons in *P. conadenia* was directly caused by the absence of the *ndh*B gene, which indirectly led to the clustering into a separate branch. In previous studies, it was found that functional genes missing in chloroplasts were transferred to the nucleus, such as the *inf* A gene in Arabidopsis, *Lotus* [39], and *Elaeagnus* [40], and the *rpl*22 gene in *Castanea* and *Passiflora* [41]. *ndh*B belongs to NADH dehydrogenase and is involved in electron transfer and oxidative phosphate processes. We investigated globally that there is no *ndh*B deletion in any other cherries. Why is the *ndh*B gene only missing in *P. conadenia*? Is it a functional gene transfer or substitution similar to the above case that occurred? At present, we are relatively poorly characterized about the biology of *P. conadenia* with this narrow distribution and further research is needed.

Previous studies have shown that the clustering hierarchical diagram based on SCUB cannot accurately reflect the real phylogenetic relationship. Compared with our phylogenetic tree based on CDS sequences, the cluster tree only reflected some interspecific relationships and provided limited supports for genetic relationships of subg. *Cerasus*. We speculate that the conflict between the topology of the clustering tree and the phylogenetic tree is due to the omission of low to medium preference codons. The CDS-based phylogenetic tree is more affected by base substitutions because it includes some non-preferred codons that also play an important role in phylogenetic relationships and would otherwise fail to reconstruct reliable genetic relationship of species.

5. Conclusions

In this study, we systematically compared the patterns of synonymous codon usage in the subg. *Cerasus* chloroplast genomes, which were all affected by natural selection rather than mutation, and the gene expression level and GC1 also contributed to some extent. In addition, clustering analysis based on RSCU values was not optimal, for which we reconstructed a reliable phylogeny based on CDS and speculated that the former was caused by the absence of the no-preference codons. Overall, we provide a new insight into the evolution of subg. *Cerasus* chloroplast genome in terms of SCUB, which will help to further explore molecular evolutionary mechanisms and optimize the expression levels of exogenous genes in the future.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/f13111891/s1, Table S1: Genbank accession numbers of 36 subg. *Cerasus* species; Table S2: RSCU value of 36 subg. *Cerasus* species; Table S3: The optimal codons of 36 subg. *Cerasus* chloroplast genome.

Author Contributions: Conceptualization, Y.-F.S. and M.L.; methodology, Y.-F.S.; software, Y.-F.S.; validation, Y.-F.S. and Q.-H.Y. and M.L.; formal analysis, Y.-F.S.; investigation, Y.-F.S. and Q.-H.Y.; resources, X.-G.Y. and Z.-Q.Z.; data curation, Y.-F.S.; writing—original draft preparation, Y.-F.S.; writing—review and editing, M.L.; visualization, Y.-F.S.; supervision, M.L.; project administration,

M.L. and X.-R.W.; funding acquisition, X.-R.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by grants from Key Modern Agriculture Project of Science and Technology Department of Jiangsu Province, China (Grant No. BE2020343).

Data Availability Statement: Not applicable.

Acknowledgments: We appreciate the assistance of Cheng Zhang (Nanjing Forestry University) during the experiments and data analyses.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Buhr, F.; Jha, S.; Thommen, M.; Mittelstaet, J.; Kutz, F.; Schwalbe, H.; Rodnina, M.V.; Komar, A.A. Synonymous Codons Direct Cotranslational Folding toward Different Protein Conformations. *Mol. Cell* **2016**, *61*, 341–351. [CrossRef] [PubMed]
2. Zhou, Z.; Dang, Y.; Zhou, M.; Li, L.; Yu, C.-H.; Fu, J.; Chen, S.; Liu, Y. Codon usage is an important determinant of gene expression levels largely through its effects on transcription. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, E6117–E6125. [CrossRef] [PubMed]
3. Long, S.; Yao, H.; Wu, Q.; Li, G. Analysis of compositional bias and codon usage pattern of the coding sequence in Banna virus genome. *Virus Res.* **2018**, *258*, 68–72. [CrossRef]
4. Wang, H.; Meng, T.; Wei, W. Analysis of synonymous codon usage bias in helicase gene from Autographa californica multiple nucleopolyhedrovirus. *Genes Genom.* **2018**, *40*, 767–780. [CrossRef] [PubMed]
5. Romero, H.; Zavala, A.; Musto, H. Codon usage in Chlamydia trachomatis is the result of strand-specific mutational biases and a complex pattern of selective forces. *Nucleic Acids Res.* **2000**, *28*, 2084–2090. [CrossRef]
6. Hunt, R.C.; Simhadri, V.L.; Iandoli, M.; Sauna, Z.E.; Kimchi-Sarfaty, C. Exposing synonymous mutations. *Trends Genet.* **2014**, *30*, 308–321. [CrossRef]
7. Pop, C.; Rouskin, S.; Ingolia, N.T.; Han, L.; Phizicky, E.M.; Weissman, J.S.; Koller, D. Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. *Mol. Syst. Biol.* **2014**, *10*, 770. [CrossRef]
8. Quax, T.E.F.; Claassens, N.J.; Söll, D.; van der Oost, J. Codon Bias as a Means to Fine-Tune Gene Expression. *Mol. Cell* **2015**, *59*, 149–161. [CrossRef]
9. López, J.L.; Lozano, M.J.; Lagares, A.; Fabre, M.L.; Draghi, W.O.; Del Papa, M.F.; Pistorio, M.; Becker, A.; Wibberg, D.; Schlüter, A.; et al. Codon Usage Heterogeneity in the Multipartite Prokaryote Genome: Selection-Based Coding Bias Associated with Gene Location, Expression Level, and Ancestry. *mBio* **2019**, *10*, e00505-19. [CrossRef]
10. Daniell, H.; Lin, C.-S.; Yu, M.; Chang, W.-J. Chloroplast genomes: Diversity, evolution, and applications in genetic engineering. *Genome Biol.* **2016**, *17*, 134. [CrossRef]
11. Kwak, S.-Y.; Lew, T.T.S.; Sweeney, C.J.; Koman, V.B.; Wong, M.H.; Bohmert-Tatarev, K.; Snell, K.D.; Seo, J.S.; Chua, N.-H.; Strano, M.S. Chloroplast-selective gene delivery and expression in planta using chitosan-complexed single-walled carbon nanotube carriers. *Nat. Nanotechnol.* **2019**, *14*, 447–455. [CrossRef] [PubMed]
12. Hishamuddin, M.S.; Lee, S.Y.; Ng, W.L.; Ramlee, S.I.; Lamasudin, D.U.; Mohamed, R. Comparison of eight complete chloroplast genomes of the endangered *Aquilaria* tree species (Thymelaeaceae) and their phylogenetic relationships. *Sci. Rep.* **2020**, *10*, 13034. [CrossRef]
13. Wang, Z.; Xu, B.; Li, B.; Zhou, Q.; Wang, G.; Jiang, X.; Wang, C.; Xu, Z. Comparative analysis of codon usage patterns in chloroplast genomes of six Euphorbiaceae species. *PeerJ* **2020**, *8*, e8251. [CrossRef]
14. Wu, Z.; Liao, R.; Yang, T.; Dong, X.; Lan, D.; Qin, R.; Liu, H. Analysis of six chloroplast genomes provides insight into the evolution of *Chrysosplenium* (Saxifragaceae). *BMC Genom.* **2020**, *21*, 621. [CrossRef] [PubMed]
15. Koehne, E. *Prunus* L. In *Plantae Wilsonianae*; Sargent, C.R., Ed.; Dioscorides Press: Portland, OR, USA, 1913; Volume 2, pp. 196–282.
16. Rehder, A. *Manual of Cultivated Trees and Shrubs Hardy in North America Exclusive of the Subtropical and Warmer temperate Regions*; MacMillan: New York, NY, USA, 1940.
17. Krüssmann, G. Cultivated broad-leaved trees and shrubs. In *Timber Press*; Timber Press: Portland, OR, USA, 1986; Volume 3.
18. Lu, L.T.; Ku, T.C.; Li, C.L.; Chen, S.X. Rosaceae (3) Prunoideae. In *Flora Reipublicae Popularis Sinicae, Tomus 38*; Yü, T.T., Ed.; Science Press: Beijing, China, 1986; pp. 1–133.
19. Wang, X.R. *An Illustrated Monograph of Cherry Cultivars in China*; Science Press: Beijing, China, 2014.
20. Zhang, Y.; Nie, X.; Jia, X.; Ding, C.; Biradar, S.; Le, W.; Che, X.; Song, W. Analysis of codon usage patterns of the chloroplast genomes in the Poaceae family. *Aust. J. Bot.* **2012**, *60*, 461–470. [CrossRef]
21. Ji, K.; Song, X.; Chen, C.; Li, G.; Xie, S. Codon Usage Profiling of Chloroplast Genome in Magnoliaceae. *J. Agric. Sci. Technol.* **2020**, *22*, 52–62.
22. Nie, X.; Deng, P.; Feng, K.; Liu, P.; Du, X.; You, F.M.; Song, W. Comparative analysis of codon usage patterns in chloroplast genomes of the Asteraceae family. *Plant Mol. Biol. Rep.* **2014**, *32*, 828–840. [CrossRef]
23. Liu, H.; Xiong, R.; Ni, Y.; Wei, L.; Sun, J.; Wang, G.; Zhang, Y.; Gao, Y. Comparative Analysis of Codon Usage Patterns in Chloroplast Genomes of *Fragaria* Species. *Mol. Plant Breed.* **2021**, 1–23.

24. Shi, S.; Li, J.; Sun, J.; Yu, J.; Zhou, S. Phylogeny and Classification of *Prunus sensu lato* (Rosaceae). *J. Integr. Plant Biol.* **2013**, *55*, 1069–1079. [CrossRef]
25. Zhang, S.-D.; Jin, J.J.; Chen, S.Y.; Chase, M.W.; Soltis, D.E.; Li, H.-T.; Yang, J.-B.; Li, D.-Z.; Yi, T.-S. Diversification of Rosaceae since the Late Cretaceous based on plastid phylogenomics. *New Phytol.* **2017**, *214*, 1355–1367. [CrossRef]
26. Jin, J.-J.; Yu, W.-B.; Yang, J.-B.; Song, Y.; dePamphilis, C.W.; Yi, T.-S.; Li, D.-Z. GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol.* **2020**, *21*, 241. [CrossRef] [PubMed]
27. Tillich, M.; Lehwark, P.; Pellizzer, T.; Ulbricht-Jones, E.S.; Fischer, A.; Bock, R.; Greiner, S. GeSeq—Versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* **2017**, *45*, W6–W11. [CrossRef] [PubMed]
28. Zhang, D.; Gao, F.; Jakovlić, I.; Zou, H.; Zhang, J.; Li, W.X.; Wang, G.T. PhyloSuite: An integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. *Mol. Ecol. Resour.* **2020**, *20*, 348–355. [CrossRef] [PubMed]
29. Sueoka, N. Near Homogeneity of PR2-Bias Fingerprints in the Human Genome and Their Implications in Phylogenetic Analyses. *J. Mol. Evol.* **2001**, *53*, 469–476. [CrossRef] [PubMed]
30. Wang, H.-X.; Liu, H.; Moore, M.J.; Landrein, S.; Liu, B.; Zhu, Z.-X.; Wang, H.-F. Plastid phylogenomic insights into the evolution of the Caprifoliaceae *s.l.* (Dipsacales). *Mol. Phylogenet. Evol.* **2020**, *142*, 106641. [CrossRef]
31. Liu, Q.; Xue, Q. Comparative studies on codon usage pattern of chloroplasts and their host nuclear genes in four plant species. *J. Genet.* **2005**, *84*, 55–62. [CrossRef]
32. Suzuki, R.; Shimodaira, H. Pvclust: An R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **2006**, *22*, 1540–1542. [CrossRef]
33. Nguyen, L.-T.; Schmidt, H.A.; Von Haeseler, A.; Minh, B.Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* **2015**, *32*, 268–274. [CrossRef]
34. Tuller, T.; Waldman, Y.Y.; Kupiec, M.; Ruppin, E. Translation efficiency is determined by both codon bias and folding energy. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 3645. [CrossRef]
35. Gu, W.; Zhou, T.; Ma, J.; Sun, X.; Lu, Z. The relationship between synonymous codon usage and protein structure in *Escherichia coli* and *Homo sapiens*. *Biosystems* **2004**, *73*, 89–97. [CrossRef]
36. Wang, B.; Yuan, J.; Liu, J.; Jin, L.; Chen, J.-Q. Codon Usage Bias and Determining Forces in Green Plant Mitochondrial Genomes. *J. Integr. Plant Biol.* **2011**, *53*, 324–334. [CrossRef] [PubMed]
37. Tang, D.; Wei, F.; Cai, Z.; Wei, Y.; Khan, A.; Miao, J.; Wei, K. Analysis of codon usage bias and evolution in the chloroplast genome of *Mesona chinensis* Benth. *Dev. Genes Evol.* **2021**, *231*, 1–9. [CrossRef] [PubMed]
38. Hershberg, R.; Petrov, D.A. Selection on codon bias. *Annu. Rev. Genet.* **2008**, *42*, 287–299. [CrossRef] [PubMed]
39. Jansen, R.K.; Cai, Z.; Raubeson, L.A.; Daniell, H.; Depamphilis, C.W.; Leebens-Mack, J.; Müller, K.F.; Guisinger-Bellian, M.; Haberle, R.C.; Hansen, A.K.; et al. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 19369–19374. [CrossRef] [PubMed]
40. Choi, K.S.; Son, O.G.; Park, S. The Chloroplast Genome of *Elaeagnus macrophylla* and *trn*H Duplication Event in Elaeagnaceae. *PLoS ONE* **2015**, *10*, e0138727. [CrossRef]
41. Jansen, R.K.; Saski, C.; Lee, S.-B.; Hansen, A.K.; Daniell, H. Complete Plastid Genome Sequences of Three Rosids (*Castanea*, *Prunus*, *Theobroma*): Evidence for At Least Two Independent Transfers of *rpl*22 to the Nucleus. *Mol. Biol. Evol.* **2011**, *28*, 835–847. [CrossRef]