Comparative chloroplast genomics of Caryophyllaceae species: Insights into sequence variations and phylogenetic evolution

LUCUN YANG (yanglucun@nwipb.cas.cn)
Chinese Academy of Sciences
Yongqing Zhu (452433958@qq.com)
Maqin County Forestry and Grassland Station
Qing Hua (1339064895@qq.com)
Golog Tibtean Autonomous Prefecture Agriculture and animal husbandry comprehensive service center

Research Article

Keywords:

DOI: https://doi.org/

Additional Declarations: No competing interests reported.

Comparative chloroplast genomics of Caryophyllaceae species: Insights into sequence variations and phylogenetic evolution

Lucun Yang[1*], Yongqing Zhu[2], Qing Hua[3]

[1]Qinghai Province Key Laboratory of Qinghai-Tibet Plateau Biological Resources, Northwest Institute of Plateau Biology, Chinese Academy of Sciences, Xining 810008, China.

[2]Maqin County Forestry and Grassland Station, Maqin 814000, China.

[3]Golog Tibetan Autonomous Prefecture Agriculture and animal husbandry comprehensive service center, Maqin 814000, China.

[*]Correspondence: Lucun Yang yanglucun@nwipb.cas.cn

Abstract

Background Caryophyllaceae contains 100 genera and 3000 species, many of which are valuable both ecologically and economically. However, as past research has shown, the fundamental phylogenetic relationships of Caryophyllaceae are still debatable, and molecular dating based on chloroplast genomes has not been thoroughly examined for the entire family.

Methods In this study, we used four newly generated Caryophyllaceae chloroplast genomes and eighteen other published genomes to clarify their genetic properties.

Results These 22 chloroplast genomes had typical quadripartite structure, with 129-134 distinct genes and lengths ranging from 133,621 bp to 153,957 bp. The 22 Caryophyllaceae chloroplast genomes showed significant variations in the number of long repeats and SSR types; mononucleotide repeats (A/T) and palindromic repeats were the most common types. Three substantially divergent areas containing *atpB-rbcL*, *rbcL-accD*, and *accD* were found by further comparative study, which could serve as effective molecular markers. The codon bias of chloroplast genomes in Caryophyllaceae were mainly affected by natural selection, but other factors such as mutation pressure could also affect the codon bias to some extent. Fourteen optimal codons were identified in the chloroplast genome of Caryophyllidae. Phylogenetic analysis demonstrated that the monophyly of any of the three recognized subfamilies within Caryophyllaceae was not supported by our data. Meanwhile, seven well-supported clades correspond to 8 tribes were found in Phylogenetic trees. The results of molecular dating demonstrated that the divergence between Caryophyllaceae and Amaranthaceae was estimated to occur in 69 Ma. Tr. Paronychieae was the oldest tribe of the eight tribes included in this study, diverged at 59.92 Ma.

Conclusion This study provides resources for further investigations on the identification, genetic engineering, evolution, and phylogeny of Caryophyllaceae species.

Keywords: Caryophyllaceae; Codon bias; Divergence hotspots; Divergence time

Introduction

As one of the largest family in angiosperm families, Caryophyllaceae Juss is made up of 100 genera and 3000 species [1], the majority of which are annual or perennial herbs or subshrubs that grow in alpine meadows, sandy grasslands, stony hillsides, fixed dunes, under coniferous forests, riverbanks, grasslands, etc [2]. It distributes in worldwide, primarily in the temperate and warm temperate parts of the Northern Hemisphere, with diversification centers in the Mediterranean Sea and the Iran-Tunisian region. With a total of 30 genera and over 390 species, the Caryophyllaceae family is primarily distributed in the north and west of China [3]. Despite having a large number of species, the Caryophyllaceae has a limited fossil record [4,5]. Simple pollen fossils appear in Australia and New Zealand about 73 million years ago in the Late Cretaceous Campanian, which was the earliest known fossil record of Caryophyllaceae [6,7]. Seed fossils were first found in England in the Eocene. [8] Studies on the biogeographic origin and distribution pattern of Caryophyllidae have been confined to the Australia taxa, which diverged considerably in the middle and late Eocene, with most extant genera arriving in Australia in the Neogene or Quaternary [9].

The plants in Caryophyllaceae are employed in many ways because of its wide diversity and adaptability. Numerous Caryophyllaceae species are highly valuable medicinal; their main chemical constituents are saponins and volatile oils. *Mesostemma gypsophiloides*, *Pseudostellaria heterophylla*, *Dianthus superbus*, *Vaccaria segetalis*, *Psammosilene tunicoides*, *Stellaria dichotoma var. lanceolata* and other species are frequently used as constituents in traditional Chinese medicine [3]. Furthermore, a large number of Caryophyllaceae species have grown to be valuable floral resources in landscaping due to their exceptional qualities, which include exquisite flower color and leaf shape as well as their high horticultural attractive value. *Lychnis*, *Dianthus*, *Silene*, *Gypsophila*, and *Saponaria* are a few examples [10]. Certain species, like *Gymnocapos przewalskii*, are first-class national protected wild plants [11]. Consequently, Caryophyllaceae species have been receiving an increasing amount of attention. However, the origin and the classification of Caryophyllaceae has been controversial.

Based on morphological characteristics, Bittrich (1993b) [12] separated the Caryophyllaceae into Alsinoideae, Caryophylloideae, and Paronychioideae subfamilies. Molecular data [13-17] demonstrated that the conventionally recognized subfamilies were non-monophyletic, however, did not support the partition of these three subfamilies. Then, in accordance with molecular phylogenetics, Harbaugh et al. (2010) [16] proposed a new classification for this family, which was backed by Greenberg & Donoghue (2011) [17]. The new classification divided the family into 11 tribes. Even though Caryophyllaceae molecular phylogenetic research has advanced in the ways mentioned above, some of the studies' findings have been inconsistent because of the small number of species included and the markers that were chosen. For instance, Greenberg and Donoghue (2011) [17] suggested that Caryophylloideae was a non-monophyletic branch, which was replaced by Tr. Eremogoneae, Tr. Sileneae and Tr. Caryophylleae, and Tr. Eremogoneae and Tr. Caryophylleae formed a sister-group relationship, which was inconsistent with Harbaugh's results. Furthermore, the range and monophyletic status of some big genera are still up for debate, and Tr. Sperguleae has a low support rate for monophyletic group. As a result, additional techniques were applied to further refine the classification system of Caryophyllaceae.

A phylogenomic framework is provided by recent developments in molecular genomics and bioinformatics, notably next-generation sequencing techniques, to map the variety and evolution of angiosperms [18-20]. The chloroplast genome differs from the nuclear genome in several ways, including maternal inheritance, excellent conservation, and suitable polymorphism. Due to these characteristics, plastome genetic polymorphism is a good source of molecular markers for a variety of genetic and phylogenetic investigations in angiosperms at various taxonomic levels []. Over the past three decades, it has become increasingly clear that modern phylogenetic analyses utilizing complete plastid genomes have significantly advanced our understanding of the links in plant evolutionary history [21]. Caryophyllaceae has been the subject of little genetic research despite its therapeutic benefits. There is currently little knowledge about Caryophyllaceae in relation to the genetic features of the chloroplast genomes. In addition, comparing the chloroplast genome of closely related species holds great potential for understanding the conservation of species and their evolutionary histories [22-25]. In this study, we sequenced the whole chloroplast genomes of four speices (*Arenaria kansuensis*, *A. roborowskii*, *A. przewalskii* and *Silene aprica*) in Caryophyllaceae. And then, we compared and analyzed these four species with other sixteen species which reported before. The primary goals of this study were to: (1) investigate the properties and genetic variations of the chloroplast genome; (2) elucidate the adaptive evolutionary of the Caryophyllaceae genomes; (3) look into the region of divergence hotspots for the purpose of differentiating the Caryophyllaceae species; and (4) reconstruct phylogenetic relationships and molecular divergence within the major lineages of Caryophyllaceae species.

Results

General features of the Caryophyllaceae chloroplast genomes

Following de novo sequencing and assembly, the four Caryophyllaceae species' complete chloroplast genomes, measuring 133,621 bp for *A. kansuensis*, 132,576 bp for *A. roborowskii*, 144, 726 bp for *A. przewalskii*, and 149,948 bp for *S. aprica*, were obtained. A small single copy region (SSC), a large single copy region (LSC), and two inverted repeat regions (IRa and IRb) are the components of the typical quadripartite structure seen in these genomes (Fig.1). A total of 22 species from 18 species of 18 genera (genome sequences are available from NCBI) and 4 newly sequenced species of Caryophyllidae were used for comparative genomic analysis. The length of the complete chloroplast genomes of all 22 Caryophyllaceae species ranged from 133,621 bp (*A. kansuensis*) to 153,957 bp (*Psammosilene tunicoides*) (Fig.2A). The lengths of the LSC, SSC, and IR regions are as follows: 74,107 bp (*Eremogone acicularis*) to 84,980 bp (*A. kansuensis*), 12,914 bp (*Lychnis wilfordii*) to 18,196 bp (*A. kansuensis*), and 20,775 bp (*A. kansuensis*) to 27,709 bp (*L. wilfordii*), respectively (Fig.2A). The IR regions have a higher GC content (40.51-44.15%) than the SSC (29.28-31.20%) and LSC (33.98-35.34%) regions (Fig.2B).

Based on gene annotation, 129-134 genes were found, including 83-89 protein-coding genes, 37-38 transfer RNAs (tRNAs), and 8 ribosomal RNAs (rRNAs) (Table 1, Table S1). There were some minor variations among these 22 chloroplast genomes, despite the fact that the majority of the protein-coding genes, tRNAs, and rRNAs were comparable. For instance, be different from *A. przewalskii*, which only had two copies of the *rpl23* gene, and *accD* and *ycf15* were absent, the chloroplast genome of *Myosoton aquaticum* had four copies of the *rpl23* gene, two copies of *ycf15*, and one copy of *accD* (Tables 1 and S1). Twenty-one of these genes—ten tRNA genes (two *trnA-UGC*, *trnG-UCC*, two *trnI-GAU*, *trnK-UUU*, *trnL-UAA*, *trnV-UAC* and two *trnH-GUG*) and eleven coding genes (*rpoC1*, two *ndhB*, *ndhA*, *petB*, *atpF*, *petD*, *rpl16*, *rps16*, and two *rps12*) contained two exons. Three exons were present in four coding genes (two each for *rps12*, *clpP1*, and *paf1*)(Table 2).Three groups of these genes were distinguished: a total of 43 genes are involved in photosynthesis (photosystem I, II, cytochrome b/f complex, ATP synthase, Rubisco large subunit, and NADPH dehydrogenase), 59 genes are related to self-replication (the large subunit of the ribosome, the small subunit of the ribosome, and RNA polymerase), and other genes are related to related enzymes (ATP-dependent protease, Maturase, Acetyl-CoA carboxylase, Cytochrome c biogenesis, and Inner membrane protein)(Table 2).

GView produced the graphical map of circular genomes to evaluate sequence differences across the 22 chloroplast genomes in Caryophyllaceae (Fig.3). The LSC and SSC region sequences in every plastome that was studied showed significant variation. The two IR regions' sequences were less diverged than the LSC and SSC regions', according to the genome comparison. Compared to the coding areas, the intergenic regions showed more divergence.

Identification of SSRs and long repetitive sequences

Microsatellites, also known as simple sequence repeats (SSRs), are widely distributed in the genomes, and are utilized as genetic markers because they are highly polymorphic, specific, and informative. They are composed of short DNA motifs, typically 1-6 bp in length. In this work, we analyzed the distribution and frequency of SSRs in 22 Caryophyllaceae chloroplast genomes. The result showed that 1,159 SSRs were found, ranged from 24 (*E. acicularis*) to 100 (

Shivparvatia glandulige) (Fig.4A). In Table S2, the precise frequency of SSRs with various repeat motifs and numbers is displayed. Of the 1,159 SSRs in total, 1057 (91.20%) were simple repeat motifs, and 102 (8.80%) were present in compound formation. Dinucleotide (p2) repeats only accounted for 3.02% of the SSRs, while mononucleotide (p1) repeats represented the largest proportion at 84.11%. At 0.09% and 0.17%, respectively, the pentanucleotide (p5) and hexanucleotide (p6) repeats were relatively rare (Table S2).

The lengths of the SSRs varied from 10 to 60 bp, with the majority falling between 10 and 15 bp (86.45%), followed by 60+ bp (4.75%), 15-20 bp (4.31%), 30-60 bp (2.50%), and 20-30bp (1.98%) (Fig.4B; Table S3). In the chloroplast genomes of *S. glanduligera*, the most abundant SSRs in 10-15 bp as well as a wide range of all sizes from 15 to 60 bp were found. In contrast, the least abundant SSRs in 10-15 bp as well as a wide range of all sizes from 15 to 60 bp were detected in *E. acicularis* (Table S3).

Moreover, the SSRs in the 22 Caryophyllaceae chloroplast genomes were more frequently located in the LSC region (70.45%) than in the SSC region (18.29%), and only a minority (5.53%) was dispersed within the IR regions (Fig.4C; Table S4). Likewise, SSRs (61.00%) in these chloroplast genomes primarily occurred in the intergenic spacer (IGS) regions, with a small portion (28.30%) distributed in CDS, while only a few (10.70%) of SSRs was found in introns regions (Fig.4D; Table S5).

The 22 chloroplast genomes of Caryophyllaceae contained 832 long repetitive sequences in total (Fig.5, Table S6). These sequences included 5-61 forward (F) repeats, 0-6 reverse (R) repeats, 0-2 complementary (C) repeats, and 7-38 palindromic (P) repeats. Palindromic (P) and forward (F) repeats made up the majority of the four different types of long repeats, with percentages of 52.40% and 42.67%, respectively, while complementary (C) and reverse (R) repeats made up just 3.37% and 1.20%, respectively.

Codon Bias in Chloroplast Genome of Caryophyllaceae

Base composition of codons

Base composition analysis was performed on the coding sequence of Caryophyllaceae chloroplast genome (Fig.6). The distribution range of GC1 (GC content of the first codon base), GC2 (GC content of the second codon base) and GC3 (GC content of the third codon base) ranged from 21.74% - 62.2%, 13.04%-56.58% and 15.62%-66.67%, respectively. The distribution frequency of GC content in the three positions of the codon is different, and the average value is GC1 (45.54%) > GC2 (39.30%) > GC3 (28.28%). Among them, GCall (total GC content of codon) is 37.71%, which is not much different from GC2. The average value of GC3 is the smallest, the selection pressure is the largest, and the A/U bias is obvious.

Analysis of the synonymous codon relative usage (RSCU) of the whole Caryophyllaceae chloroplast genome (Fig.7) showed that the Caryophyllaceae coding sequence contained 64 types of codons. Among them, thirty-one of the chloroplast genome codons have RSCU≥1(Table S7), of which 29 end in A/U, making up 97%, demonstrating a clear A/U bias.

Neutrality-plot analysis

Fig.8 showed that there was very little association between GC_{12} and GC_3, with a regression coefficient of 0.227 and a correlation coefficient of 0.291(R^2=0.085). Natural selection was the primary factor influencing the codon preference of the Caryophyllaceae chloroplast genome, as evidenced by the fact that most of the genes of the Caryophyllaceae chloroplast coding sequence were located above the diagonal line, with only a few genes being close to or below the line.

ENC-plot analysis

Fig.9 showed that more genes were distributed below and away from the expected curve and fewer genes were distributed on the expected curve. This suggests that natural selection, rather than mutation pressure, is the primary factor affecting the use bias of the chloroplast genome codon in Caryophyllaceae, with the majority of the genes' actual ENC values differing from their theoretical ENC values.

PR2-plot analysis

The codon bias analysis of chloroplast genome of Caryophyllaceae is shown in Fig.10. The scatters of the four regions in the PR2 plan are not evenly distributed. The majority of genes are found near the bottom (< 0.5) of the G_3/GC_3 axis, with a small number at the top (> 0.5). The majority of genes are found on the left (<0.5) of the A_3 / AU_3 axis, while a small number are found on the right (>0.5). This suggests that G > C and A > T occurrences exist at the third position of the synonymous codon of the four nucleotides. Given that mutation pressure is the only factor influencing codon use bias, the distribution of synonymous codons, C and G and A and T, should be identical on the third position. Therefore, natural selection as well as mutation have an impact on the codon use bias of the chloroplast genome of Caryophyllaceae.

Determination of the optimal codon

Table 3 showed that there were 16 codons that satisfied the requirements RSCU>1 and △RSCU⩾ 0.08 concurrently. Therefore, these 16 codons (AAU, UGU, CAA, GAA, CAU, UAU, GGU, CCU, ACA, GUU, AGA, CGA, CUU, UUG, AGU, UCA) were identified as the optimal codons of Caryophyllaceae chloroplast genome, of which 6 end in A and 9 end in U. The results showed that Caryophyllaceae chloroplast genome preferred to use A/U ending codons, which was consistent with the results of GC_3 and RSCU analysis. Therefore, when using Caryophyllaceae chloroplast gene engineering to design exogenous gene vectors, selecting codons ending in A/U can improve the expression and transformation efficiency of exogenous genes.

IR contraction and expansion

To identify distinctive and shared characteristics, the border regions of the LSC, SSC, and IR regions of the 22 Caryophyllaceae cp. genomes were examined (Fig.11). These chloroplast genomes showed generally stable patterns with comparable gene richness and organization with the exception of the *L. wilfordii* and *A. przewalskii*. The LSC/IRb boundary was located within the *rps*19 gene (with the 3ˊ end of the *rps*19 located in the LSC region while 5ˊ end located in the IRb), with spanned 59-180 bp in LSC region and 21-220 bp in IRb region. In both *L. wilfordii* and *A. przewalskii*, *rps*19 gene were lost in the LSC/IRb boundary, and *rpl2* gene was transferred from IRb region to LSC region. The shortened copy of *ycf1* gene spanned the IRb/SSC border and interlaced with the *ndhF* gene. The shortened copy of *ycf1* gene was mostly found in the IRb region, with one end extending from 0 bp (*M. dichotomum*) to 96 bp (*P. argentea*) into the SSC region. On the other hand, the majority of *ndhF* gene was found in the SSC region, where it partially overlapped with the duplicated ycf1 gene. And the length of the section found in the IRb region varied from 2 bp in *Paronychia argentea* to 66 bp in *Psammosilene tunicoides* and *Gymnocarpos przewalskii*. The shortened copies of *ycf1* gene were missing in both *L. wilfordii* and *A. przewalskii*, and the *ndhF* and *pbf1* genes were indented to the SSC region by 100bp and 81bp, respectively. The SSC/ IRa junction was located in the *ycf1* coding region, with a size variation from 3,380 bp (*S. glanduligera*) to 3,882 bp (*P. argentea*). At the SSC/IRa border, the *ycf1* gene extended into the SSC region, at varying lengths ranging from 1,761 bp in *P. argentea* to 1, 921 bp in *Stellaria neglecta*. The SSC/ IRa junction of *L. wilfordii* was located within the *rps15* gene, and the distance between *rps15* and SSC/IRa border was 62 bp, while the SSC/ IRa junction of *A. przewalskii* was located within the *ndhA* gene, and with its end extending 10bp into the SSC region. The IRa/LSC border was located within *trnH* gene, but was located 0 bp (*P. missionariorum*) to 39 bp (*Stellaria neglecta* and *Pseudostellaria davidii*) apart from the IRa/LSC border.

Genome comparison and sequence divergence analyses

We used mVISTA to identify the divergent regions in the multiple alignments of 22 Caryophyllaceae chloroplast genomes (Fig.12). Higher degree variants were found mostly in the IGS regions, such as, *rps16-trnG-UCC*, *ycf1-trnR-ACG*, *ndhF-rp132*, *ycf2-trnL-CAA*, *ndhF-rpl32*, *atpB-rbcL*, *atpF-atpH*, *atpH-atpI*, t*rnE-UUC-trnT-GGU*, *psbE-petL*, and *psaC-ndhE*. Additional variants were found in the intron-containing genes, including *rps16*, *petD*, *atpF*, *rpoC1*, r*pl16*, and *ycf1*. Apart from a few genes with sequence variants, like *atpI*, *rbcL*, *psaI*, *accD*, *clpP1*, *ycf2*, *ndhF*, *ycf3* and *ndhA*, the majority of the genes in the CDS area were found to be reasonably well conserved. The rRNA genes of these species, however, showed a significant degree of conservation.

Using DnaSP software, the nucleotide variability (Pi) value was found in order to evaluate the degree of sequence divergence in the chloroplast genomes of the 22 Caryophyllaceae species. With a mean of 0.059051, the Pi values of the 22 species ranged from 0.00177 to 0.21727 (Fig.13). The IR regions showed lower levels of nucleotide polymorphisms than the LSC and SSC regions. Furthermore, Pi values (>0.1877) were exceptionally high in 10 divergent locations, all of which were located in the LSC (Table S8). Among them, seven divergent regions (*trnF-GAA*, *trnF-GAA_ndhJ*, *ndhC_trnM-CAU*, *trnM-CAU*, *trnM-CAU_atpE*, *atpB_rbcL*, *rbcL_accD*) were located in noncoding intergenic regions, and three (*atpE, atpB, accD*) was within protein-coding regions, (Table S8). Such regions of high variation can serve as potential markers for species authentication and population genetics analysis in this family.

Phylogenetic relationships

As seen in Fig.14, ML analyses of the whole chloroplast genomes supported the monophyletic of Caryophyllaceae. The first divergence within Caryophyllaceae separates a clade comprised of Gymnocarpos and Paronychia (the tribe Paronychieae of Harbaugh & al., 2010) from the rest of Caryophyllaceae (100% BS; node A, Fig.14). The first divergence within node B diverges into the final clade of Paronychiodeae included in this study (designated as tribe Sperguleae by Harbaugh & al., 2010) and the rest of Caryophyllaceae (100% BS; node b, Fig.14). The first divergence within node C divides a clade of Alsinoideae species (the tribe Sclerantheae of Harbaugh & al., 2010) from the rest of Caryophyllaceae (100% BS; node C, Fig.14). The first divergence within node D separates another clade of Alsinoideae (designated as tribes Arenariean and Alsineae by Harbaugh & al., 2010) from the rest of Caryophyllaceae (100%BS; node D, Fig.14). The first divergence within node E divides a clade of Caryophylloideae species (designated as tribe Caryophylleae by Harbaugh & al., 2010) from the rest of

Caryophyllaceae (100% BS; node C, Fig.14). The large remaining Caryophyllaceae clade (100% BS; node F, Fig.14) comprises other members of subfamilies Alsinoideae and Caryophylloideae, and is split into two large clades (100% BS and 100% BS, respectively; nodes G and F, Fig.14), which corresponds respectively to tribes Eremogoneae and Sileneae in Harbaugh's study.

Divergence Time Estimation of Caryophyllaceae

In this study, the divergence times of the major clades in the Caryophyllaceae were estimated using the complete chloroplast genome sequences of eighty species, representing eighteen genera, eight tribes, as well as two outgroups. The divergence between Caryophyllaceae and Amaranthaceae was estimated to occur in 69 Ma (million years) (Fig.15). Tr. Paronychieae was the oldest tribe of the eight tribes included in this study, diverged at 59.92 Ma. Tr. Sperguleae and other 6 tribes approximately diverged in 47.18 Ma. Tr. Sileneae was the most evolved clades of Caryophyllaceae, it diverged with Tr. Eremogoneae probably at 34.66 Ma. The estimated divergence time in 80 species of Caryophyllaceae was between 26.47 and 0.54 Ma.

Discussion

Plastid genome features

The usual quadripartite structure (one LSC region, one SSC region, and two IR regions) that has been reported in other angiosperms species was also observed in 22 complete chloroplast genomes of Caryophyllaceae in this study [26-28]. In these 22 chloroplast genomes, gene loss and duplication occurred despite the great degree of conservation observed in the majority of the protein-coding genes, tRNAs and rRNAs. For examples, *L. wilfordii* lost *ycf15* and *accD* and had only two copies of *rpl23* in its chloroplast genome and *A. przewalskii* had two copies of *trnQ*-UUG only in its chloroplast genome, indicating that *L. wilfordii* and *A. przewalskii* underwent gene loss and insertion during their evolutionary processes. On the contrary, in other chloroplast genomes of higher plants, reports of other gene loss and duplication had been made. For example, *ndh* genes had been lost in the families Gentiaceae [29], Orobanchaceae [30] and Orchidaceae [26], and *trnS-GCU* and *trnT-UGU* had been duplicated in *Globba schomburgkii* [31]. The gene content of the IR borders across Caryophyllaceae plastomes was similar, and the IR regions were generally more conservative than the LSC and SSC regions. Still, minor differences in the border locations between the IR and SC regions were found. The ycf1 gene crossed the IRa/SSC boundary regions in all species, resulting in a pseudogene—an incomplete duplication or shortened copy—of this gene inside IRs. The *ycf*1 pseudogene overlapped with the *ndhF* gene at the IRb/SSC junction in each of these cp. genomes, resulting in different fragment lengths at the IRb region. Previous research has demonstrated a primary correlation between the stability of the IR/SC boundary regions and the transformation of gene *ndhF* and/or *ycf*1[26, 32-34]. We found that the IR/SC boundaries displayed minor fluctuations across Caryophyllaceae species. These changes were mainly associated with the different positions of *ndhF* and *ycf1*, together with the genes *rps19* and *trnH* adjacent to LSC/IR and SSC/IR borders.

Repeat sequence analysis

The 22 Caryophyllaceae plastid genomes showed an unequal distribution of polymorphic SSRs, with differences in the quantity, size, and kind of SSR motifs, according to repetitive sequence analysis. Similarly, these genomes' lengthy repetitive sections showed a different distribution of repeat types. The emergence of distinct motifs for various SSR types may be the consequence of selecting pressures. According to Carmona et al. [35], variations in the distribution and quantity of repetitive DNA sequences are important factors that propel speciation and genome evolution. In addition, SSRs have been employed as molecular markers to examine population genetics and polymorphisms, as well as to detect notable degrees of variation in closely related species. Therefore, these non-overlapping sequence repeats and SSRs can all be utilized to make markers for genetic diversity studies of various Caryophyllaceae species.

Codon Bias in Chloroplast Genome of Caryophyllaceae

Different species exhibit non-random distribution of synonymous codons, leading to codon preference. An essential metric for examining the evolutionary relationships between the chloroplast genome in plants is codon preference. Additionally, different species or even different genes within the same species may exhibit distinct codon bias. Naturally selection and mutation pressure are the main determinants of codon use preference [36]. The use preference of the codon is closely related to the GC content of the codon. Because the third position of the codon is less affected by selection pressure, GC3 is usually used as an important parameter for the analysis of codon usage bias. In this study, the codon GC content of Caryophyllaceae chloroplast genome was less than 50%, indicating that Caryophyllaceae chloroplasts are more inclined to use A/T codons. The claim made by Campbell and Gowri [37] that "higher plant codons tend to use A/T endings" is further supported by the low GC content of the Caryophyllaceae chloroplast genome codon GC3.

Neutrality-plot and ENC-plot analysis of the Caryophyllaceae chloroplast genome showed that natural selection had a greater influence on the chloroplast genome's codon usage bias than mutation pressure does. PR2-plot analysis of the Caryophyllidae chloroplast genome revealed that natural selection as well as mutations had an impact on the chloroplast genome's codon usage bias. Although natural selection and mutational pressure can both produce codon use preference on their own, the primary factor in the formation of codon use preference for Caryophyllidae is the interaction of these two processes and their long-term cumulative effect [38]. This finding is consistent with the chloroplast genomes of *Panicum miliaceum* [36], *Betula alnoides*[39], and *Mangifera indicate* [40]. However, natural selection is the primary factor influencing the preference of codon use in the research of

Camellia oleifera [41] and *Gynostemma pentaphyllum* [42], whereas mutation has a little effect. These findings suggest that the variables influencing the chloroplast genome's codon bias vary amongst plants.

In the chloroplast genome of Caryophyllaceae, there are 16 codons of protein-coding genes (AAU, UGU, CAA, GAA, CAU, UAU, GGU, CCU, ACA, GUU, AGA, CGA, CUU, UUG, AGU, and UCA) that simultaneously match the requirements RSCU > 1 and ΔRSCU≥ 0.08. These codons are identified as the best codons in the chloroplast genome of Caryophyllidae, with the exception of one that ends in G, all the others ending in A and U. This suggests that the use of codons in Caryophyllaceae tends to the third codon position of A and U, and has strong A/U base preference. Similar findings were obtained by *Bothriochloa ischaemum* [43], 29 Magnoliaceae plants [44], and *Tribulus terrestris*[45]. These findings suggest that most plants have a substantially conserved chloroplast genome codon use pattern.

Comparative genomes

Comparative analysis showed that the LSC and SSC regions of 22 chloroplast genomes of Caryophyllaceae were found to be more diverged than the IR regions, which is in line with findings for other plants [27-28, 46]. Previous phylogenetic analyses of Caryophyllaceae using 3 chloroplast fragments (*matK*, *trnL-F* and *rps*16) and 5 chloroplast fragments (*matK*, *ndhF*, *trnL-F*, *trnQ-rpsl6* and *trnS-trnf*) have yielded inconsistent results [16-17]. It was also evident from the Pi values examined in this work that the commonly employed chloroplast genome markers, such as *matK*, *ndhF* and *rps*16, had relatively modest polymorphisms (0.073, 0.095 and 0.051, respectively) at the tribe level. Three divergent hotspot regions (*atpB-rbcL*, *rbcL-accD*, and *accD*) among the 22 whole chloroplast genomes of Caryophyllaceae have been found based on Pi values in this study. These variable areas may thus be appropriate as prospective DNA markers for Caryophyllaceae species identification and phylogenetic relationships research.

Phylogenetic relationship and divergence time of

Caryophyllaceae

Although morphological characteristics have historically led to the division of the Caryophyllaceae into three major subfamilies—Alsinoideae, Caryophylloideae, and Paronychioideae [12,47]—it has not been evident how much molecular data supports or refutes these divisions [13-16]. Harbaugh et al. (2010) [16], however, proposed a different tribal categorization for the group based on evidence of the non-monophyly of at least the Paronychioideae. The monophyly of any of the three recognized subfamilies within Caryophyllaceae is not supported by our data. Our findings, however, closely align with those of Harbaugh et al (2010) [16]. Our findings place Eremogoneae, a tiny clade that includes Arenaria subg. Eremogone and subg. Eremogoneastrum, as a sister group to Sileneae, which includes Sliene and Arenaria przewalskii. Meanwhile, subfamilies Alsinoideae and Caryophylloideae form a clade together. As a result, neither the classic Caryophyllodieae nor the Alsinoideae are monophyletic. Meanwhile, subfamily Paronychioideae is a non-monophyletic grade of early diverging lineages. In addition, our findings mostly agree with the tribal classification of Harbaugh et al. (2010) [16], while it is challenging to make direct comparisons because we have included a few numbers of taxa. We also cannot exactly define the limits of these taxa since phylogenetic definitions [48] are still pending. All of the tribes identified by Harbaugh et al. (2010) [16] are supported as monophyletic by our tree, with very few exceptions. Our phylogeny shows that Caryophylloideae is a non-monophyletic branch, which is replaced by the tribes Eremogoneae, Sileneae and Caryophylleae, and the tribes Sileneae and Eremogoneae form a sister group relationship, which is inconsistent with the finding of Harbaugh [16] and Greenberg [17]. Additionally, in the phylogenetic tree, tribes Alsineae and Arenariean form a clade, indicating that these two tribes are not monophyletic. Moreover, our findings, in fact, supported the suggestions put forth by Harbaugh et al. (2010) [16] and Greenberg et al. (2011)[17] regarding the phylogenetic position of Arenaria species based on their phylogenetic results and physical traits such grass-like leaves, suggesting that the Arenaria species in this clade belong to a new tribe called Eremogoneae.

Previous studies have shown that simple pollen fossils of Caryophyllaceae appeared in Australia and New Zealand about 73 Ma ago during the Late Cretaceous Campanian, which is the earliest known fossil record of Caryophyllaceae [49-50]. Seed fossils first appeared in Britain during the Eocene Epoch [51]. In this study, the divergence between Caryophyllaceae and Amaranthaceae was estimated to occur in 69 Ma, which was similar to simple pollen fossils (73 Ma). In addition, previous studies have suggested that the ancestral range of the tribe Alsineae was reconstructed into Central Asia, so the divergence of the tribe Alsineae may be related to the uplift of the Tibetan Plateau. Our findings supported the results put forth by Zhang [52] regarding the differentiation time of tribe Alsineae (25.87 Ma).

Seven tribes that currently proposed classification systems for Caryophyllaceae were better supported by our findings. However, for the whole Caryophylliaceae, the use of only 81 genome sequences is far from sufficient. Consequently, to better solve the phylogenetic relationships within Caryophylliaceae and provide a crucial foundation for the study of the biogeographic evolution of Caryophylliaceae, future research must integrate the taxa that are challenging to sample and combine the chloroplast genome data, especially the genera and species that have never been sampled.

Conclusion

In the chloroplast genomes of 22 Caryophylliaceae species, we identified the genomic characteristics, sequence divergences, and mutation patterns in this study. Genome differences between genera and species were identified through comparison of genomic sequences, which also offered important insights into the overall evolutionary dynamics of the Caryophylliaceae. A strong backbone phylogeny of Caryophyllaceae with well-resolved deep nodes was produced by our phylogenomic analyses. The findings show that the relationships between the major groups are strongly supported, but they also show that some tribes are not monophyly. Future research that includes a large taxonomic sample as well as morphological evidence is therefore required.

Methods

Plant material and sampling

In the wild in Qinghai Province, fresh young leaves of four distinct species (*Arenaria kansuensis* Maxim (GSXLZ), *A. roborowskii* Maxim (QZXLZ), *Silene aprica* Turcz. ex Fisch. et Mey. (NLC), and *A. przewalskii* Maxim) were sampled. The locations where the four plants were sampled were as follows: Qumalai County (95.2010′E, 34.6720′N, 4600 m), Mengyuan County (101°22′47.55′E, 37°20′23.42′ N, 4010 m), Maqin County (101°24′0.6″E, 34°27′38″ N, 3,538 m), and Maqin County (102.22′E, 37.45′N, 3,400 m), respectively. Using silica gel, the leaves were quickly preserved until they dried. Prof. Yuhu Wu, a taxonomist at the Northwest Institute of Plateau Biology, Chinese Academy of Sciences, identified each of the samples. These four species' voucher specimens were placed under the following voucher numbers: QHGC20230821, QHGC20230829, QHGC20230911, and QHGC20230915, respectively, at the Qinghai-Tibetan Plateau Museum of Biology (QTPMB). From GenBank, all complete chloroplast genomes of Caryophyllaceae that have been published were retrieved. 81 accessions from 80 species of 18 genera were retrieved in total (Table S9). Institutional, governmental, and international rules are followed in all aspects of our experimental study, including the gathering of plant samples.

DNA extraction, Sequencing, Assembly, and Annotation

Using a G-spin™II for Plant Genomic DNA extraction kit (iNtRON, Seoul, Korea), the young leaf's total genomic DNA was extracted. Using electrophoresis on a 1% Tris-acetate (TAE)–ethylenediamine tetra acetic acid (EDPA) agarose gel, the purity and quality of the DNA were assessed.

Following the isolation of genomic DNA, 5-10 µg of DNA was sheared, and then adapter ligation and library amplification were carried out. Shanghai Peisenor Biotechnology Co., LTD. [Shanghai, China] sequenced the raw pair-end reads using Illumina NovaSeq technology. To trim Illumina raw reads, NGSQCToolkitv2.3.3's Trimming function was utilized [53]. Using the cp genome of the closely related species *E. acicularis* (NC_069855) as a reference [54], clean reads were assembled using MIRA v4.0.2 after low-quality reads and adapters were removed. Then, MITObim v1.8 was used to further assemble the desired contigs [55].

Using the contigs that were acquired, GeneiousR8 v8.0.2 (Biomatters Ltd., Auckland, New Zealand) produced a consensus sequence [56]. The Dual Organellar Genome Annotator programme (DOGMA) was used to annotate the entire cp genome. In Geneious R8 v8.0.2, the start and stop codons were manually adjusted for gene annotation based on the annotation of other cp genomes. Additionally, tRNA scan SE1.21 was used to confirm each and every tRNA gene. The MAUVE programme was used to align sequences in order to compare the genomes' structure and gene contents. The circular complete chloroplast genome map for every species was created using Organellar Genome DRAW v1.1 (OGDRAW) (http://ogdraw.mpimp-golm.mpg.de)[57]. Four Caryophyllaceae species' recently discovered cp genomes have been deposited in the Gene Bank with corresponding accession numbers (OR863397-OR863400).

Codon Bias analyses

Codon composition analysis

CodonW 1.4.2 was used to analyze coding sequences of Caryophyllaceae chloroplast genome, and the relative usage (RSCU) and effective codon number (ECN) of each CDS sequence were obtained [58] (Sharp and Li,1987). GC content (GC1, GC2, GC3) and average GC content (GCall) at three codon locations were analyzed using online software (CUSP) (http://emboss.toulouse.inra.fr /cgi-bin/emboss/cusp). SPSS and EXCEL software were used to analyze the results.

ENC is often used to evaluate the degree of synonym codon use bias, and its value ranges from 20 to 61. ENC value 45 is the cut-off point. The smaller the value, the stronger the bias, and the larger the value, the weaker the bias. RSCU is the ratio of the actual frequency of a codon to the theoretical frequency. RSCU = 1, indicating that the codon does not use bias; RSCU > 1 indicates that the codon is used more frequently than expected, and vice versa indicates that the codon occurs less frequently than other synonymous codons [59].

Neutrality-plot analysis

Analysis of the variables influencing codon use bias is done using neutral plots. Each dot in the picture represents a gene; the vertical coordinate is the GC_{12} content (the average value of GC_1 and GC_2), and the horizontal coordinate is the GC_3 content. The codon choice is mostly influenced by mutation pressure if the regression coefficient is near to 1 and all of the scatter points in the figure are spread diagonally. This suggests that the codon's base composition is identical. Conversely, it suggests that selection pressure has a significant impact on its preference [60].

ENC−plot analysis

ENC-plot plots include standard curves and scatter plots. Scatter plots take ENC and GC_3 as vertical and horizontal coordinates, respectively. The formula of the standard curve is $ENC=2+GC_3+29/\ [GC_3{}^2+\ (1-GC_3)\ {}^2]$, which means that when there is no selection pressure, the nucleic acid sequence of the gene determines the codon preference. The specific criterion is the distance between the scatter point and the standard curve in the figure. If the distance between the two is closer, the main influencing factor is the base composition, and the other is the selection pressure [61].

PR2▯plot analysis

Using PR2-plot analysis, the variables influencing nucleotide composition were identified. The horizontal and vertical coordinates of the plot were A_3/

(A_3+U_3) and G_3/ (G_3+C_3), respectively. The center point of the graph represents A=T, C=G, which means that the codon bias is not affected by selection pressure, and the vector distance between the remaining points and the center point indicates the direction and degree of its bias [62].

Determination of the optimal codon

The ENC values of the gene sequences obtained after the Caryophyllaceae chloroplast genome screening were sequenced from high to low, and 10% genes were selected from both ends of the lowest and highest values to construct the high-low expression database. The RSCU values and ΔRSCU (the difference between the high-low expression databases) were computed using CodonW 1.4.2. The codon satisfying ΔRSCU≥0.08 and RSCU > 1 is identified as the optimal codon [63].

Repeats and SSR analyses

The programmer REPuter v.2.74 [64] (https://bibiserv.cebitec.unibielefeld.de/reputer/) was used to examine palindrome repeats and scattered repeats in Caryophyllaceae plastomes, including forward, reverse, and complement repeat sequences. The following conditions were applied in order to identify these oligonucleotide repeats: a hamming distance of 3 (i.e., 90% or higher sequence identity); a minimum repeat size of 30 bp. Furthermore, using a Perl script-based programmer called MISA v.1.01, the genomes' microsatellites and simple sequence repeats (SSRs) were analyzed [65]. A predetermined minimum threshold of 10, 5, 4, 3, 3, and 3 repeat units was used to calculate the various lengths of SSRs for mono-, di-, tri-, tetra-, penta-, and hexa-nucleotides, respectively.

Plastome comparison and sequence divergence analyses

Using 100 bp connection windows, BLAST Atlas on the GView server (https://server.gview.ca/) was utilized to visualize and evaluate the characteristics of the chloroplast genome [66]. The IRscope web application was used to study and compare the expansion and shrinkage of the IR regions of various chloroplast genomes [67]. Using mVISTA v.2.0's Shuffle-

LAGAN mode, the diverging regions were plotted [68-69]. Nucleotide diversity (Pi) values were calculated by DnaSP v6.12.03 software [70] with a sliding window analysis. The window length was set to 600 bp with a step size of 200 bp.

Phylogenetic analyses

To deduce the phylogenetic relationships within Caryophyllaceae, we performed a phylogenetic analyse using maximum likelihood (ML) method based on complete plastome sequences. A total of 81 accessions from 80 species of 18 genera of 8 tribes (Tr. Paronychieae, Tr. Sperguleae, Tr. Alsineae, Tr. Arenariean, Tr. Caryophylleae, Tr. Sclerantheae, Tr. Sileneae and Tr. Eremogoneae) representing the main lineages of Caryophyllaceae were contained, plus two outgroup species (*Amaranthus tricolor* (NC_065013) and *Cyathula officinalis* (OP936078)). We were unable to obtain the material of Tr. Polycarpeae, Tr. Corrigioleae, and Tr. Sagineae, which were not included in the analyses. Using MAFFT v7.313, all genome sequences were aligned [71], and BioEdit was used to make manual adjustments [72]. The ML tree was generated using FastTree 2[73] and implemented in RAxML v.8.2.11 [74] under the generalized time-reversible GTR + G model. Nodes were evaluated by Shimodaira–Hasegawa (SH) tests [75] to detect significant topology.

Divergence time estimation

To calculate the divergence times of Caryophyllaceae species, BEAST v1.8.4 was used [76]. The investigation comprised the sequences of the chloroplast genomes from 80 species belonging to the Caryophyllaceae as well as outgroups. Phylosuite can be used to convert the sequence alignment result file into nex format. BEAUti in BEAST v1.8.4 can be used to define the site model's parameters. The optimal Model GTR is generated by the Phylosuite v1.2.1 program's Model Finder plug-in. Next, choose the Relaxed clock log Normal as the model for the molecular clock, and leave the parameters at their default settings. Pollen fossils of Caryophyllaceae from Campanian sediments in in Australia and New Zealand was used as lognormal priors, with an offset at 73 Ma [77], a mean of 0.7, and a standard deviation of 1.0. For a duration of 2 × 10^7 generations, the Markov Chain Monte Carlo (MCMC) chains were utilized, sampling every 2000 generations and discarding the first 25% of warmed trees as burn-in. The xml file is created and executed using BEAST v1.8.4 once all the parameters have been configured. After running the log file, look at the Tracer distribution diagram and effective sample size (ESS) in Tracer v1.7 [78]. Adjust the MCMC algebra such that the ESS value is larger than 200, indicating that the running parameters have converged, if the ESS value is less than 200. Maximum clade credibility (MCC) trees were generated with TreeAnnotator v2.4.1, using a 10% burn-in (as trees), a 0.5 posterior probability limit, and a median height for node selection [79]. The time tree was edited and visualized using FigTree v1.4.4 [80].

Author’s contributions

Lucun Yang: Methodology, Software, Investigation, Writing – original draft, Writing – review & editing. **Yongqing Zhu:** Software, Investigation. **Qing Hua**: Investigation.

Funding

This research was financially supported by the Qinghai Basic research Program [Grant No. 2023-ZJ-775].

Data availability

All the newly sequenced sequences in this study have been submitted to the NCBI database (https://www.ncbi.nlm.nih.gov/genbank/) with GenBank accession numbers shown in Table S9 (OR863397-OR863400). Submitted data will remain private until related manuscript has been accepted. All data generated or analyzed are included within the article and the supplementary information files.

Declarations

Ethics approval and consent to participate

This study including the collection of plant samples complies with relevant institutional, national, and international guidelines and legislation. All the necessary permissions have been granted for this research.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

References

1. Hernández-Ledesma P, Berendsohn WG, Borsch T, Mering SV, Akhani H, Arias S, Castañeda-Noa I, Eggli U, Eriksson R, Flores-Olvera H, Fuentes-Bazán S, Kadereit G, Klak C, Korotkova N, Nyffeler R, Ocampo G, Ochoterena H, Oxelman B, Rabeler RK, Sanchez A, Schlumpberger BO & Uotila P. A taxonomic backbone for the global synthesis of species diversity in the angiosperm order Caryophyllales. Willdenowia. 2015; 45: 281–383.
2. Bittrich V. The Families and Genera of Vascular Plants. In: Kubitzki K, Rohwer J and Bittrich V, Eds., Flowering Plants: Dicotyledons; Magnoliid, Hamamelid and Caryophyllid Families, Springer, Berlin, 1993; 206-236.
3. Dequan L and Morton J. Flora of China. In: Wu, Z. and Raven, P.H., Eds., Caryophyllaceae through Lardizabalaceae, Science Press, Beijing and Missouri Botanical Garden Press, St. Louis, 2001;1-113.
4. Nowicke, JW. Caryophyllales: Evolution and Systematics. In: Behnke, HD and Mabry, JT, Eds., Pollen Morphology and Exine Ultrastructure, Springer Verlag, Berlin, 1994; 168-221.
5. Punt, W and Hoen, PP. Caryophyllaceae. Rev Palaeobot Palyno.1995; 88:82-272.
6. Stover, LE and Partridge, AD. Tertiary and Late Cretaceous Spores and Pollen from the Gippsland Basin, South-Eastern Australia. Proc R Soc Vic. 1973; 85, 237-286.
7. Raine, JI. Outline of a Palynological Zonation of Cretaceous to Paleogene Terrestrial Sediments in West Coast Region, South Island, New Zealand. New Zealand Geological Survey Report.1984; 109:1-82.
8. Chandler, EM. Flora of the Lower Headon Beds of Hampshire and the Isle of Wight. Bulletin of the British Museum of Natural History: Geology, 1961; 5:91-158.
9. Gregory, JJ and Michael KM. A Middle-Late Eocene inflorescence of Caryophyllaceae from Tasmania, Australia. Am J Bot. 2003; 90:761-768.
10. Lu DQ. Caryophyllaceae in Loess Plateau from China. Acta Bot Boreal-Occident Sin. 1994; 14(5):121-127.
11. Fu, LG. China Plant Red Data Book. Science Press, Beijing, 1999; 202.
12. Bittrich, V. Caryophyllaceae. Pp. 206–236 in: Kubitzki, J. (ed.), The families and genera of vascular plants, Berlin: Springer, 1993b.
13. Smissen, RD, Clement, JC, Garnock-Jones, PJ & Chambers, GK. Subfamilial relationships within Caryophyllaceae as inferred from 5′

ndhF sequences. Amer J Bot. 2002; 89: 1336–1341.

14. Fior S, Karis PO, Casazza G, Minuto L & Sala F. Molecular phylogeny of the Caryophyllaceae (Caryophyllales) inferred from chloroplast *matK* and nuclear rDNA ITS sequences. Amer J Bot. 2006; 93: 399–411.
15. Frajman B, Eggens F & Oxelman B. Hybrid origins and homoploid reticulate evolution within Heliosperma (*Sileneae,Caryophyllaceae*): A multigene phylogenetic approach with relative dating. Syst Biol. 2009; 58: 328–345.
16. Harbaugh DT, Nepokroeff M, Rabeler RK, McNeill J, Zimmer EA & Wagner WL. A new lineage-based tribal classification of the family Caryophyllaceae. Int J Pl Sci. 2010; 171:185–198.
17. Greenberg AK & Donoghue MJ. Molecular systematics and character evolution in Caryophyllaceae. Taxon. 2011; 60 (6): 1637-1652.
18. Ravi V, Khurana JP, Tyagi AK, Khurana P. An update on chloroplast genomes. Plant Syst Evol. 2008;271(1–2):101–22.
19. Yang JB, Tang M, Li HT, Zhang ZR, Li DZ. Complete chloroplast genome of the genus Cymbidium: lights into the species identification, phylogenetic implications and population genetic analyses. BMC Evol Biol. 2013; 13:84.
20. Daniell H, Lin CS, Yu M, Chang WJ. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biol. 2016; 17:134.
21. Gitzendanner MA, Soltis PS, Yi TS, Li DZ, Soltis DE. Plastome phylogenetics: 30 years of inferences into plant evolution. In: Chaw SM, Jansen RK, editors. Advances in botanical research. Volume 85. Cambridge: Academic Press; 2018; 293–313
22. Liu, HB, Lu, YZ, Lan, BL Xu JC. Codon usage by chloroplast gene is bias in *Hemiptelea davidii*. J Genet. 2020; 99, 8.
23. Moore MJ, Bell CD, Soltis PS, Soltis DE. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. Proc Natl Acad Sci USA. 2007;104:19363—8.
24. Moore MJ, Soltis PS, Bell CD, Burleigh JG, Soltis DE. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. Proc Natl Acad Sci USA. 2010;107:4623—8.
25. Cui Y, Chen X, Nie L, Sun W, Hu H, Lin Y, Li H, Zheng X, Song J, Yao H. Comparison and phylogenetic analysis of chloroplast genomes of three medicinal and edible Amomum species. Int J Mol Sci. 2019; 20: 4040.
26. Dong WL, Wang RN, Zhang NY, Fan WB, Fang MF, Li ZH. Molecular evolution of chloroplast genomes of orchid species: insights into phylogenetic relationship and adaptive evolution. Int J Mol Sci. 2018;19(3):716.
27. Li DM, Liu HL, Pan YG, Yu B, Huang D, Zhu GF. Comparative Chloroplast Genomics of 21 Species in Zingiberales with Implications for Their Phylogenetic Relationships and Molecular Dating. Int J Mol Sci. 2023;

24:15031.

28. Xiong C, Huang Y, Li ZL, Wu L, Liu ZG, Zhu WJ, Li JH, Xu R and Hong X. Comparative chloroplast genomics reveals the phylogeny and the adaptive evolution of *Begonia* in China. BMC Genomics. 2023; 24:648.
29. Ebert D, Peakall R. Chloroplast simple sequence repeats (cpSSRs): Technical resources and recommendations for expanding cpSSR discovery and applications to a wide array of plant. Mol Ecol Resour. 2009;9:673—90.
30. Frailey DC, Chaluvadi SR, Vaughn JN, Coatney CG, Bennetzen JL. Gene loss and genome rearrangement in the plastids of five Hemiparasites in the family Orobanchaceae. BMC Plant Biol. 2018;18:30.
31. Li DM, Li J, Wang DR, Xu YC, Zhu GF. Molecular evolution of chloroplast genomes in subfamily Zingiberoideae (Zingiberaceae). BMC Plant Biol. 2021; 21:558.
32. Kim KJ, Lee HL. Complete chloroplast genome sequences from korean ginseng (Panax schinseng Nees) and comparative analysis of sequence evolution among 17 vascular plants. DNA Res. 2004; 11:247–61.
33. Luo J, Hou BW, Niu ZT, Liu W, Xue QY, Ding XY. Comparative chloroplast genomes of photosynthetic orchids: insights into evolution of the Orchidaceae and development of molecular markers for phylogenetic applications. PLoS ONE. 2014;9(6):e99016.
34. Kim HT, Kim JS, Moore MJ, Neubig KM, Williams NH, Whitten WM, Kim JH. Seven new complete plastome sequences reveal rampant independent loss of the ndh gene family across orchids and associated instability of the inverted repeat/small single-copy region boundaries. PLoS ONE. 2015;10(11): e0142215.
35. Carmona A, Friero E, de Bustos A, Jouve N, Cuadrado A. Cytogenetic diversity of SSR motifs within and between Hordeum species carrying the H genome: H. Vulgare L. and H. Bulbosum L. Theor Appl Genet. 2013; 126:949–61.
36. Li G, Zhang L, Xue P. Codon usage pattern and genetic diversity in chloroplast genomes of *Panicum* species. Gene. 2021;15(802):145866
37. Campbell WH, Gowri G. Codon usage in higher plants, green algae, and cyanobacteria. Plant Physiol. 1990;92(1):1-11.
38. Hong SR, Lin SL, Li YP, Li YY, Li HY, Zhang QB. Analysis of the complete chloroplast genome sequence characteristics and its code usage bias of *Sorghum bicolor*. Acta Agrestia Sinica. 2023; 31(12):3636-3650.
39. Yuan XL, Li YQ, Wang Y, and Zhang JF. Analysis of codon usage in the chloroplast genome of *Betula alnoides*, Jiyinzuxue yu Yingyong Shengwuxue. Genomics and Applied Biology. 2020; 39(12): 5758-5764.
40. Tang YJ, Zhao Y, Huang GD, Fu HT, Song EL, Li RW, Jin G. Analysis on Codon Usage Bias of Chloroplast Genes from Mango. Chinese Journal of Tropical Crops. 2021; 42(8): 2143-2150.
41. Wang PL, Yang LP, Wu HY, Nong YL, Wu SC, Xiao YF, Qin ZH, Wang HY,

Liu HL. Condon preference of chloroplast genome in *Camellia oleifera*. Guihaia. 2018; 38(2) :135-144.

42. Zhao YM, Yang GQ, Zhang X, Luo ZH, Wang Q, Ding B. Codon Usage Bias of Chloroplast Genome in *Gynostemma pentaphyllum*. Molecular Plant Breeding, 2021
43. Gao SY, Li YY, Yang ZQ, Dong KH, Xia FS. Codon usage bias analysis of the chloroplast genome of *Bothriochloa ischaemum*. Acta Prataculturae Sinica. 2023; 32(7): 85-95.
44. Ji KK, Song XQ, Chen CG, Li G, Xie SQ. Codon Usage Profiling of Chloroplast Genome in Magnoliaceae. Journal of Agricultural Science and Technology. 2020; 22(11): 52-62.
45. Yang GF, Su KL, Zhao YR, Song ZB, Sun J. Analysis of codon usage in the chloroplast genome of *Medicago truncatula*. Acta Prataculturae Sinica. 2015; 24(12):171-179.
46. Li L, Wu QP, Zhai JW, Wu KL, Fang L, Li MZ, Zeng SJ and Li SJ. Comparative chloroplast genomics of 24 species shed light on the genome evolution and phylogeny of subtribe Coelogyninae (Orchidaceae). BMC Plant Biol. 2024; 24:31
47. Chrtek, J & Slavikova, Z Leitbündelanordnung in den Kronblättern von ausgewählten Arten der Familie Stellariaceae. Preslia. 1987; 60: 11-21.
48. Cantino PD, Doyle JA, Graham SW, Judd WS, Olmstead RG, Soltis DE, Soltis PS, and Donoghue MJ. Towards a phylogenetic nomenclature of Tracheophyta. Taxon. 2007; 56: 822- 846.
49. Stover, LE and Partridge, AD. Tertiary and Late Cretaceous Spores and Pollen from the Gippsland Basin, South-Eastern Australia. Proceedings of the Royal Society of Victoria.1973; 85:237-286.
50. Raine JI. Outline of a Palynological Zonation of Cretaceous to Paleogene Terrestrial Sediments in West Coast Region, South Island, New Zealand. New Zealand Geological Survey Report. 1984; 109:1-82.
51. Chandler EM. Flora of the Lower Headon Beds of Hampshire and the Isle of Wight. Bulletin of the British Museum of Natural History: Geology, 1961; 5:91-158.
52. Zhang RX. Phylogeny and biogeography of the Genus Cerastium (Caryophyllaceae). Shanxi: Shanxi Normal University, 2020.
53. Patel RK, Jain M. NGS qc toolkit: A toolkit for quality control of next generation sequencing data. PLoS ONE. 2012;7: e30619.
54. Bakker FT, Lei D, Yu JY, Mohammadin S, Wei Z, van de Kerke S, Gravendeel B, Nieuwenhuis M, Staats M, Alquezar-Planas DE. Herbarium genomics: Plastome sequence assembly from a range of herbarium specimens using an iterative organelle genome assembly pipeline. Biol J Linn Soc. 2016;117:33-43.
55. Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. Using SPAdes de novo assembler. Curr Protoc Bioinforma. 2020;70(1):e102.
56. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S,

Buxton S, Cooper A, Markowitz S, Duran C. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012;28(12):1647—9.

57. Marc L, Oliver D. Sabine K and Ralph B Organellar Genome DRAW—a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. Nucleic Acids Res. 2013; 41:75-81.
58. Sharp PM & Li WH. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Research. 1987; 15(3):1281-1295.
59. Liu HB, Lu YZ, Lan BL, Xu JC. Codon usage by chloroplast gene is bias in *Hemiptelea davidii.* Journal of Genetics. 2020; 99(1):8.
60. Kawabe A and Miyashita NT. Patterns of codon usage bias in three dicot and four monocot plant species. Genes Genet Syst. 2003; 78:343-352.
61. Wright F. The ‘effective number of codons’ used in a gene. Gene.1990; 87:23-29.
62. Sueoka N. Near homogeneity of PR2-bias fingerprints in the human genome and their implications in phylogenetic analyses. J Mol Evol. 2001; 53(4-5):469-476.
63. Liu QP and Xue QZ. Codon usage in the chloroplast genome of rice (Oryza sativa L. ssp. japonica). Acta Agron Sin. 2004; 30:1220—1224.
64. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res. 2001;29(22):4633-42.
65. Beier S, Thiel T, Münch T, Scholz U, Mascher M. MISA-web: A web server for microsatellite prediction. Bioinformatics. 2017;33:2583—5.
66. Petkau A, Stuart-Edwards M, Stothard P, van Domselaar G. Interactive microbial genome visualization with GView. Bioinformatics. 2010; 26:3125–6.
67. Amiryousefi A, Hyvönen J, Poczai P. IRscope: an online program to visualize the junction sites of chloroplast genomes. Bioinformatics. 2018;34(17):3030–31.
68. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S, Program NCS. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res. 2003;13(4):721–31.
69. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. Nucleic Acids Res. 2004;32:W273–9.
70. Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos- Onsins SE, et al. DnaSP 6: DNA sequence polymorphism analysis of large data sets. Mol Biol Evol. 2017;34(12):3299–302.
71. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol.

2013, 30(4): 772-780.

72. Hall TA, BioEdit. A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser. 1999;41:95–8.

73. Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS ONE. 2010;5(3):e9490.

74. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3.

75. Shimodaira H, Hasegawa M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol. 1999;16:1114–6.

76. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior summarization in Bayesian Phylogenetics using tracer 1.7. Syst Biol. 2018;67(5):901—4.

77. Wang SM, Li J, Wu SD, Wang W, Zhang LJ. New Advances in Caryophyllaceae Systematics. Botanical Research. 2017; 6(3): 103-113.

78. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior summarization in Bayesian Phylogenetics using tracer 1.7. Syst Biol. 2018;67(5):901—4.

79. Rambaut A, Drummond AJ. TreeAnnotator version 1.6.1 [computer program]. http:// beast. bio. ed. ac. uk.

80. Rambaut, A. FigTree v.1.4.4. Available at:http:// tree. bio. ed. ac. uk/ softw are/ figtree/. (Accessed 25 Oct 2020). 2018.

Figures

Figure 10

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Tables.docx
- TableS1.xlsx
- TableS2.xlsx
- TableS5.xlsx
- TableS3.xlsx
- TableS4.xlsx
- TableS6.xlsx
- TableS7.xlsx
- TableS9.xlsx
- TableS8.xlsx