



Article

Comparative Chloroplast Genome Analysis of Wax Gourd (*Benincasa hispida*) with Three Benincaseae Species, Revealing Evolutionary Dynamic Patterns and Phylogenetic Implications

Weicai Song ¹ , Zimeng Chen ¹, Li He ², Qi Feng ¹, Hongrui Zhang ¹, Guilin Du ³ , Chao Shi ^{1,4,*} and Shuo Wang ¹

¹ College of Marine Science and Biological Engineering, Qingdao University of Science and Technology, Qingdao 266042, China; weicai1123@163.com (W.S.); chenzimeng1209@163.com (Z.C.); qi_feng107@163.com (Q.F.); zhanghongrui1785@163.com (H.Z.); shuowang@qust.edu.cn (S.W.)

² Aesthetic Education Center, Qingdao University of Science and Technology, Qingdao 266042, China; heli2612@163.com

³ Lab of Biorefinery, Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai 201210, China; dugl1@shanghaitech.edu.cn

⁴ Plant Germplasm and Genomics Center, Germplasm Bank of Wild Species in Southwest China, Kunming Institute of Botany, The Chinese Academy of Sciences, Kunming 650204, China

* Correspondence: chaoshi@qust.edu.cn

Abstract: *Benincasa hispida* (wax gourd) is an important Cucurbitaceae crop, with enormous economic and medicinal importance. Here, we report the de novo assembly and annotation of the complete chloroplast genome of wax gourd with 156,758 bp in total. The quadripartite structure of the chloroplast genome comprises a large single-copy (LSC) region with 86,538 bp and a small single-copy (SSC) region with 18,060 bp, separated by a pair of inverted repeats (IRA and IRB) with 26,080 bp each. Comparison analyses among *B. hispida* and three other species from Benincaseae presented a significant conversion regarding nucleotide content, genome structure, codon usage, synonymous and non-synonymous substitutions, putative RNA editing sites, microsatellites, and oligonucleotide repeats. The LSC and SSC regions were found to be much more varied than the IR regions through a divergent analysis of the species within Benincaseae. Notable IR contractions and expansions were observed, suggesting a difference in genome size, gene duplication and deletion, and the presence of pseudogenes. Intronic gene sequences, such as *trnR-UCU-atpA* and *atpH-atpI*, were observed as highly divergent regions. Two types of phylogenetic analysis based on the complete cp genome and 72 genes suggested sister relationships between *B. hispida* with the *Citrullus*, *Lagenaria*, and *Cucumis*. Variations and consistency with previous studies regarding phylogenetic relationships are discussed. The cp genome of *B. hispida* provides valuable genetic information for the detection of molecular markers, research on taxonomic discrepancies, and the inference of the phylogenetic relationships of Cucurbitaceae.

Keywords: *Benincasa hispida*; chloroplast genome; comparative analysis; divergence region; phylogenetic



Citation: Song, W.; Chen, Z.; He, L.; Feng, Q.; Zhang, H.; Du, G.; Shi, C.; Wang, S. Comparative Chloroplast Genome Analysis of Wax Gourd (*Benincasa hispida*) with Three Benincaseae Species, Revealing Evolutionary Dynamic Patterns and Phylogenetic Implications. *Genes* **2022**, *13*, 461. <https://doi.org/10.3390/genes13030461>

Academic Editor: Zhiqiang Wu

Received: 21 February 2022

Accepted: 1 March 2022

Published: 4 March 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Cucurbitaceae is a moderately large family of about 130 genera and 900 species [1]. Because of their economic importance in temperate regions, species of the Cucurbitaceae family have a long and close association with human beings [2]. Familiar edible and medicinal fruits, such as cucumber (*Cucumis sativus*), melon (*Cucumis melo*), watermelon (*Citrullus lanatus*), bottle gourd (*Lagenaria siceraria*), pumpkin, and squash (*Cucurbita* spp.) are the main crops of Cucurbitaceae [3]. All are economically valuable fruit crops. The early molecular phylogeny of Cucurbitaceae was reconstructed using five chloroplast markers, which weakly support two subfamilies of Cucurbitoideae and Nhandiroboideae [4]. Recent

studies have reported that the phylogenetic tree of Cucurbitaceae contains a new classification of 15 tribes and 95–97 genera, using 14 molecular markers from the nuclear, plastid, and mitochondrial genomes [5]. However, the relationships between these subfamilies are still unresolved, possibly due to limited phylogenetic signals of the molecular markers, with a large proportion (over 70%) of missing data [6]. Comprehensive and complete sequence information is a reliable foundation for phylogenetic studies of Cucurbitaceae.

Benincasa represents a monotypic genus with a single species, belonging to tribe Benincaseae (Cucurbitaceae), which is not difficult to find in the markets since: (1) it is a highly commercialized vegetable due to its long shortage properties; and (2) it bears giant fruit, normally 80 cm in length and with a weight of over 20 kg [7,8]. Wax gourds are widely distributed in temperate and sub-temperate climates, such as China, Japan, Korea, India, and several tropical countries. Currently, it is being increasingly popular in the Caribbean and the United States [8]. Wax gourd has important nutritional and medical applications as an important vegetable crop [9]. Its pharmaceutical values cover various aspects, including central nervous system diseases (muscle tension, Alzheimer's disease [10]), gastroprotective diseases [11], depression-like activities [12], diabetes, dropsy, diseases related to the liver, urinary diseases [13], and heart diseases. Other effects, including hypolipidemic, antioxidant, anti-inflammatory, antipyretic, anti-angiogenic [14] and antimicrobial properties of *B. hispida* are also reported [15–17]. Studies have reported that the seeds of *B. hispida* contain saponin, urea, citrulline, oleic acid, and fatty acids [18,19].

The chloroplast (cp) is a self-replicating organelle that consists of homogeneous circular DNA molecules. The double-strand DNA inside the cp genome ranges from 70 to 520 kb in algae and is generally more conserved in land plants, ranging from 120 to 160 kb. Although the specific nucleotide sequences vary across different species, the quadripartite structure and organization retain a firm consistency, which can be classified into four sections: a large single-copy (LSC) region and a small single-copy (SSC) region. These are separated by a pair of inverted repeats (IRa and IRb) [20]. As a metabolic center, the cp genome remains highly conservative to sustain the normal physiological function of cells, especially for genes related to photosynthesis. Despite its conservative nature, there are partial differences in gene types and genome sizes, such as substitutions, insertions, and deletions of nucleotide sites; contractions and expansions of IR regions; and rearrangements and translocations of genes [21,22]. This polymorphism and diversity can be used in population taxonomic and phylogenetic analysis, population genetics studies, and evolutionary investigations [23,24].

Although wax gourd is widely consumed among Asian communities, no detailed resources have been published regarding its genomic features, nor has a comparative analysis with its related species. To date, the nuclear gene of wax gourd has been reported [25], whereas the whole nuclear genome has some limitations in phylogenetic analysis of species. Therefore, we sequenced and assembled the complete chloroplast genome sequence of *B. hispida* and submitted the data to the National Center for Biotechnology Information (NCBI). This study is the first elaborate report of the cp genome of *B. hispida*, and the first comparative analysis to include three other species from Benincaseae, namely *Lagenaria siceraria*, *Citrullus colocynthis*, and *Citrullus lanatus*. We aimed to reveal: (1) the quadruple structure and the composition of different regions and functions; (2) putative RNA editing sites; (3) patterns of repeats and microsatellite; (4) highly divergent regions; (5) the phylogenetic relationships among Cucurbitaceae. The results provided may contribute to the unfolding of taxonomical discrepancies, identifying suitable genetic markers, and the inference of phylogenetic positions among related species.

2. Materials and Methods

2.1. Plant Material, DNA Extraction, and Sequencing

Fresh leaves of *Benincasa hispida* were collected from Panlong District, Kunming City, China (24°23' N, 102°10' E), and the voucher specimen and DNA were deposited at Qingdao University of Science and Technology (specimen code DG200618). Fresh leaf tissue was collected without apparent disease symptoms and preserved in silica gel. Total genomic

DNA was extracted from fresh leaves using modified CTAB [26], the quantity and quality of extracted DNA was assessed by spectrophotometry, and the integrity was evaluated using a 1% (*w/v*) agarose gel electrophoresis [27]. An Illumina TruSeq Library Preparation Kit (Illumina, San Diego, CA, USA) was used to prepare approximately 500 bp of paired-end libraries for DNA inserts, according to the manufacturer's protocol. These libraries were sequenced on the Illumina HiSeq 4000 platform in Novogene (Nanjing, China), generating raw data of 150 bp paired-end reads. About 14.6 Gb high quality, 2×150 bp pair-end raw reads were obtained and were used to assemble the complete chloroplast genome of *B. hispida*.

2.2. Genome Assembly and Annotations

The raw data were preprocessed using Trimmomatic 0.39 software [28], including removal of Adapter sequences and other sequences introduced in the sequencing, removal of low-quality and over-N-base reads, etc. The quality of newly produced clean short reads was assessed using FASTQC v0.11.9 and MULTIQC software [29,30], and high-quality data with Phred scores averaging above 35 were screened out. According to the reference sequence (*Cucumis melo*), the chloroplast-like (cp) reads were isolated from clean reads by BLAST [31]. Short reads were de novo assembled into long contigs with SOAPdenovo 2.04 [32] by setting kmer values as 35, 44, 71, and 101. Finally, the long-contigs complete sequence expansion and gap filling using Geneious ver. 8.1 [33], forming the complete chloroplast genome. The complete chloroplast genome was further validated and calibrated by using de novo splicing script NOVOplasty 4.2 [34]. GeSeq [35] was used to annotate the de novo assembled genomes, tRNAscan-SE ver 1.21 [36] was applied to detect tRNA genes with default settings, and RNAmmer [37] was used to validate rRNA genes with default settings. As a final check, we compared the results with the reference sequence and manually corrected the erroneous genes by GB2Sequin [38]. The circular map of the genomes was drawn by using Organellar Genome DRAW (OGDRAW) [39]. The newly assembled *B. hispida* chloroplasts genomes were deposited in GenBank, with the accession number MW362306.

2.3. Chloroplast Genome Comparison

In order to gain a better understanding of the characteristics of the cp genome of *B. hispida*, we selected three species that are not only closely related to *B. hispida* but also representative of Benincaseae to perform comparative analysis. Sequences of their complete chloroplast genome were downloaded from NCBI database, with the following accession numbers: *Lagenaria siceraria* (MT773628), *Citrullus colocynthis* (NC_035727), and *Citrullus lanatus* (KY430692).

2.4. Codon Usage and Putative RNA Editing Site

Codon usage and amino acid frequency were calculated by Geneious Prime[®] 2020 [40], and relative synonymous codon usage (RSCU) of protein-coding genes was evaluated by MEGA-X [41]. We also used predictive RNA editors for plant chloroplast (PREP-cp) [42] to investigate putative RNA editing sites in the cp genomes of *B. hispida*, *C. colocynthis*, *C. lanatus*, and *L. siceraria*.

2.5. Repeat Sequences and SSR Analysis

MicroSATellite identification tool (Misa) [43] was used to determine simple sequence repeats (SSRs) or microsatellites in cp genomes of four species. SSRs were determined by a settled minimum threshold of nine for mononucleotide repeats, four for dinucleotide, and three for tri-, tetra-, penta-, and hexanucleotide repeats. Oligonucleotide repeats were analyzed by REPuter program [44] to find four types of repeats, including forward (F), reverse (R), complementary (C), and palindromic (P). These four types of repeats were detected with a minimum repeat size of 20 bp, edit distance of 3, and 90% similarities.

2.6. Comparative Analysis of cp Genomes in Benincaseae

IRscope [45] was used to detect the contraction and expansion of IRs boundaries, which were visualized between four main regions in chloroplast genome (LSC/IRb/SSC/IRa). The mVISTA program [46] was used to compare the cp genome of four species using Shuffle-LAGAN model with *Lagenaria siceraria* set as the reference sequence.

DnaSP [46] was used to perform sliding window analysis using multiple alignment of complete cp genome of four selected species, along with the determination of synonymous (Ks) and non-synonymous (Ka) substitutions and their ratio (Ka/Ks). Geneious was used to detect the types, numbers, lengths, and positions of SNPs and InDels in LSC, SSC, and IR regions. To further evaluate their natural selection pressure, genes that presented Ka/Ks value greater than one were tested with site model using CodeML [47] algorithm implemented in EasyCodeML [48]. The likelihood ratio test (LRT) was used to compare seven codon substitution models (M0, M1a, M2a, M3, M7, M8, and M8a). The Bayes empirical Bayes (BEB) evaluated the posterior probability of positive selection sites.

2.7. Phylogenetic Analysis

We selected and downloaded the sequences of 23 species from Cucurbitales and three outgroup species including *Libidibia coriaria* (NC_026677), *Glycine max* (NC_007942) and *Solanum lycopersicum* (NC_007898) from NCBI to perform phylogenetic tree building. Maximum likelihood (ML) tree was constructed through two approaches. One phylogenetic tree was constructed using complete cp genome and the other was built with 72 gene sequences. MAFFT alignment was performed using 72 concatenated gene sequences and the best-fit model was found by MEGA-X [41]. All indels was excluded for both alignments, leaving only substitutions for ML analysis. The best-fit models applied for all three were GTR + G, determined based on Bayesian information criterion (BIC) [49].

3. Results

3.1. Chloroplast Genome Assembly, Organization, and Features of *Benincasa hispida*

The paired-end sequencing of *B. hispida* by Illumina HiSeq 4000 generated around 14.6 GB raw data with 82.6 million 150 bp reads. We de novo assembled its complete chloroplast genome and the data were submitted to NCBI under accession number MW362306 after a thorough check for correctness. As shown in Table 1 and Figure 1, the size of its complete chloroplast genome is 156,758 bp in length, presenting a typical quadripartite structure with a large single-copy region (LSC, 86,538 bp), a small single-copy region (SSC, 18,060 bp) and two inverted repeat regions (IRa/b, 26,080 bp each).

Table 1. Chloroplast genome general features of *Benincasa hispida*.

Characteristics	<i>Benincasa hispida</i>	
Size (base pair, bp)	156,758	
LSC length (bp)	86,538	
SSC length (bp)	18,060	
IR length (bp)	26,080	
Number of genes	131	
Number of protein-coding genes	86	
Number of tRNA genes	37	
Number of rRNA genes	8	
Duplicate genes	18	
GC content	Total (%)	37.2
	LSC (%)	35
	SSC (%)	31.7
	IR (%)	42.9
	CDS (%)	37.9
	rRNA (%)	55.2

Table 1. Cont.

Characteristics	<i>Benincasa hispida</i>
tRNA (%)	53.2
ALL gene %	39.4
Protein-coding part (CDS) (% bp)	51.1
All genes (% bp)	71.6
Non-coding region (% bp)	28.4

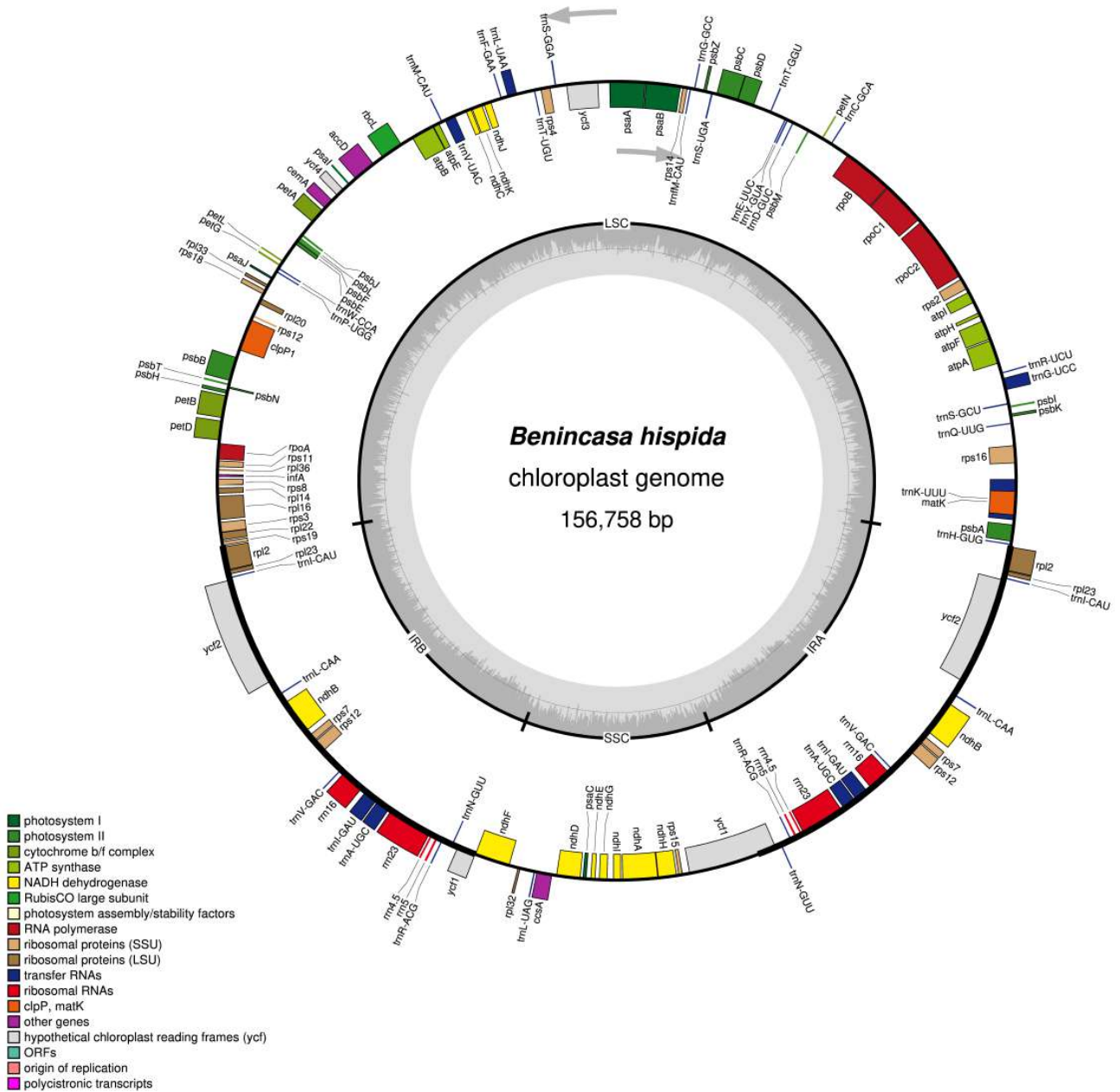


Figure 1. Gene map of the *Benincasa hispida* chloroplast genome. The genes drawn outside and inside of the circle are transcribed in clockwise and counterclockwise directions. Genes are colored based on their function. The borders of chloroplast genome are defined with LSC, SSR, IRa, and IRb. The dashed gray color of the inner circle shows the GC content, whereas the lighter gray color shows AT content. Asterisks mark genes that have introns.

The cp genome of *B. hispidula* had 131 genes (Table 2), including 86 protein-coding genes, 37 tRNA genes, and 8 rRNA genes, 18 of which were duplicated genes (7 protein-coding genes, 7 tRNA genes, and 4 rRNA genes). The total GC content of the cp genome was 37.2%, with the IR regions having the highest GC content at 42.9%, followed by LSC (35%) and SSC (31.7%). In terms of the GC contents of the different gene types, the number of rRNA (55.2%) and tRNA (53.2%) was relatively high, and that of CDS was 37.9%. In total, 18 genes contained introns, 16 of which (10 protein-coding genes and 6 tRNA genes) contained 1 intron, and 2 CDSs (*ycf3* and *clpP1*) possessed 2 introns (Table S1). Among these genes, 17 genes were duplicated in the IR regions except one trans-splicing gene, which was observed in the *rps12* gene with 5'-end located in the LSC region and 3' end duplicated in the IR regions. The truncation event was observed in the *ycf1* gene, which started in the IRa region and ended at the SSC region, leaving a 100 bp truncated copy in the IRb region.

Table 2. Genes predicted in the chloroplast genome of *Benincasa hispidula*. The number of asterisks after the gene names indicates the number of introns contained in the genes.

Category of Genes	Group of Genes	Gene Name
Photosynthesis-related genes	Large subunit of rubisco	<i>rbcL</i>
	Photosystem I	<i>psaA, psaB, psaC, psaI, psaJ</i>
	Assembly/stability of photosystem I	<i>ycf3</i> **, <i>ycf4</i>
	Photosystem II	<i>psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ</i>
	ATP synthase	<i>atpA, atpB, atpE, atpF</i> *, <i>atpH, atpI</i>
	Cytochrome b6/f complex	<i>petA, petB</i> *, <i>petD</i> *, <i>petG, petL, petN</i>
	Cytochrome c synthesis	<i>ccsA</i>
	NADH dehydrogenase	<i>ndhA</i> *, <i>ndhB</i> *, <i>ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK</i>
Transcription and translation related genes	RNA polymerase subunits/transcription	<i>rpoA, rpoB, rpoC1</i> *, <i>rpoC2</i>
	Small subunit of ribosomal proteins	<i>rps11, rps12</i> * (*2), <i>rps14, rps15, rps16</i> *, <i>rps18, rps19, rps2, rps3, rps4, rps7</i> (*2), <i>rps8</i>
	Large subunit of ribosomal proteins	<i>rpl14, rpl16</i> *, <i>rpl2</i> * (*2), <i>rpl20, rpl22, rpl23</i> (*2), <i>rpl32, rpl33, rpl36</i>
	Translation initiation factor	<i>infA</i>
RNA genes	Ribosomal RNA	<i>rrn16</i> (*2), <i>rrn23</i> (*2), <i>rrn4.5</i> (*2), <i>rrn5</i> (*2)
	transfer RNA	<i>trnA</i> -UGC * (*2), <i>trnR</i> -ACG (*2), <i>trnR</i> -UCU, <i>trnN</i> -GUU (*2), <i>trnD</i> -GUC, <i>trnC</i> -GCA, <i>trnQ</i> -UUUG, <i>trnE</i> -UUC, <i>trnG</i> -GCC, <i>trnG</i> -UCC *, <i>trnH</i> -GUG, <i>trnI</i> -CAU (*2), <i>trnI</i> -GAU * (*2), <i>trnL</i> -CAA (*2), <i>trnL</i> -UAA *, <i>trnL</i> -UAG, <i>trnK</i> -UUU *, <i>trnM</i> -CAU, <i>trnM</i> -CAU, <i>trnF</i> -GAA, <i>trnP</i> -UGG, <i>trnS</i> -GCU, <i>trnS</i> -GGA, <i>trnS</i> -UGA, <i>trnT</i> -GGU, <i>trnT</i> -UGU, <i>trnW</i> -CCA, <i>trnY</i> -GUA, <i>trnV</i> -GAC (*2), <i>trnV</i> -UAC *
Other genes	RNA processing	<i>matK</i>
	Carbon metabolism	<i>cemA</i>
	Fatty acid synthesis	<i>accD</i>
	Proteolysis	<i>clpP1</i> **
	Component of TIC complex	<i>ycf1</i> (*2)
	Hypothetical proteins	<i>ycf2</i> (*2)

* Gene with one intron, ** gene with two introns, (*2) gene with two copies.

3.2. Codon Usage and Amino Acid Frequencies

The complete cp genome of *Benincasa hispida* contained 80,109 bp of coding sequences (CDSs) that encoded 86 genes, including 26,703 codons that fit in 64 codon types. The results of the amino acid frequency analysis showed that leucine, with 10.5% occurrence, was the most abundant amino acid, followed by isoleucine, with 8.5%. Cysteine, with only 1.1% abundance, was the amino acid that occurred the least.

The relative synonymous codon usage (RSCU) of the four species was also calculated, presenting a high codon bias of A or T bases. The distribution of the codon usage showed that the codons ending with A or T had RSCU > 1 except GGT (Glycine, 0.96), AGT (Serine, 0.9), and CGT (Arginine, 0.68), revealing that the codons ending with A or T were preferred, while the codons ending with C or G were non-preferred. Among all three stop codons, TAA, with 64% abundance, was the most frequent (Table S2).

3.3. Putative RNA Editing Site within *Benincaseae*

RNA editing events are typical in the cp genomes of most land plants and essential for understanding the chloroplast genome at the transcript level. For this purpose, we determined the RNA editing site in the cp genomes of four species from *Benincaseae*. In the cp genome of *Benincasa hispida*, PREP-web found 58 putative RNA editing sites in 21 CDS (Table S3a). Among these genes, the *ndhB* gene, with thirteen editing sites, was determined to be the most variant gene, followed by *ndhD* (eight sites) and *rpoB* (five sites). We also found that 81% of all RNA editing events occurred at the second nucleotide position of the codons, while none of these events were located in the third codon position.

Moreover, these RNA editing events resulted in post-transcriptional substitutions, causing amino acid conversions. In the group of these conversions, fifty-four out of fifty-six RNA editing sites led to hydrophobic products, comprising phenylalanine (9), isoleucine (5), leucine (32), methionine (2), valine (4), and tryptophan (2). Four exceptions led to hydrophilic (neutral) amino acid products, including cysteine (1), tyrosine (2), and serine (1). Furthermore, serine-to-leucine was found to be the most abundant post-transcriptional substitution, with 41.82% of all RNA editing events, followed by proline-to-leucine (14.55%) and serine-to-phenylalanine (7.27%). It is worth mentioning that two RNA editing events were detected that transformed ACG (Thr) to the initiation codon AUG, resulting in the start of translation in the *ndhB* and *ndhD* genes.

As shown in Table S3b, the total number of RNA editing sites detected was 57 in *Citrullus lanatus* and 55 in *Lagenaria siceraria* and *Citrullus colocynthis*. All the patterns mentioned above showed high consistency in all four species analyzed, with only minor differences in terms of numerical values.

3.4. Repeated Sequence and SSR Analysis

In this study, we analyzed microsatellites or simple sequence repeats (SSRs) in the cp genome of *Benincasa hispida*, *Citrullus lanatus*, *Lagenaria siceraria*, and *Citrullus colocynthis* using MISA-web, and high similarity was revealed between the four species (Figure 2). We found that *B. hispida* contained the most abundant number of SSRs (238), while *C. lanatus*, with only 219 SSRs, had the least. In the cp genome of *B. hispida*, most of the SSRs were mononucleotides (42%), varying from 9 to 15 repeat units. Meanwhile, the abundance of dinucleotide was only 25%, which was slightly lower than that of trinucleotide (30%). The frequencies of tetranucleotide and pentanucleotide were only 3% and 0.42%, respectively, and hexanucleotide repeats were absent from all the species (Figure 2C). Moreover, most of the mononucleotide repeats were A/T motifs, while AT/TA motifs comprised 68% of the dinucleotide repeats (Table S4).

We also analyzed the distribution of SSRs in two different types of regions, specifically LSC/IR/SSC regions and intergenic spacer (IGS)/gene regions. According to the results, most of the repeats were located in the LSC region, varying from 136 in *C. lanatus* to 148 in *B. hispida*, followed by the SSC region (38 in *B. hispida*) and IR regions. Noticeably, the SSC number in the IR regions was 26 in all the species except for *L. siceraria* (24), implying that

the IR regions were more conserved than the LSC and SSC regions (Figure 2A). The IGS regions were determined to have a high abundance of SSRs in comparison with the gene regions. We found 125 SSRs within 46,150 bp IGS regions and 116 SSRs in 112,281 bp gene regions, meaning the density of SSRs in the IGS regions was 2.62 times of that of the gene regions (Figure 2B). Similar results were present in all the species.

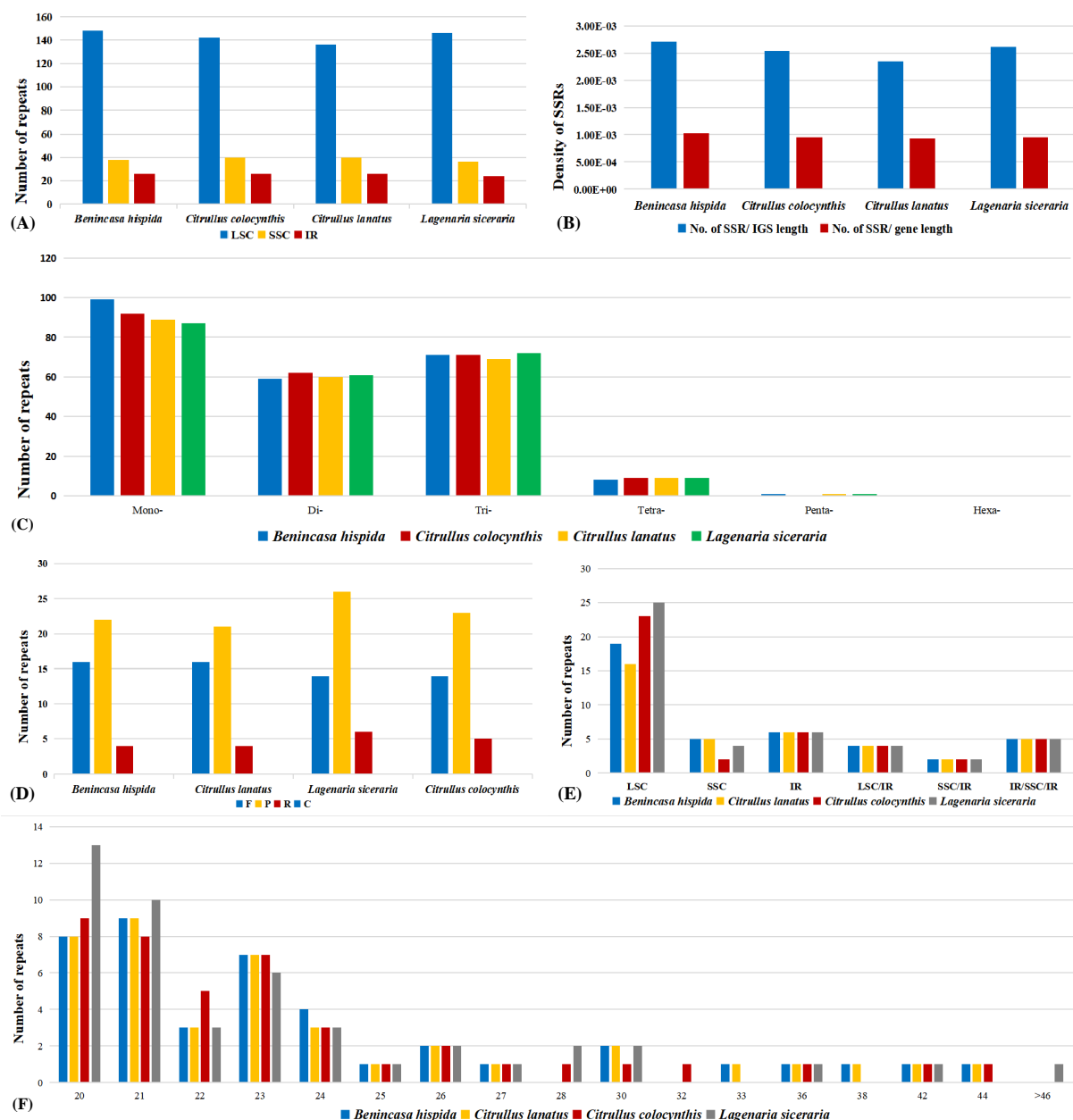


Figure 2. Comparison of microsatellites and oligonucleotide repeats in the chloroplast genomes of *Benincaseae* species. (A) The number of SSRs in the three main regions of the chloroplast genome. LSC: large single-copy region, SSC: small single-copy region, IR: inverted repeat region. (B) The density of the SSRs in the IGSs (intergenic sequences) and gene regions. (C) The number of different types of SSR. Mono- represent mononucleotide SSRs, Di- represent dinucleotide SSRs, etc. (D) Different types of oligonucleotide repeat. F: forward repeats, P: palindromic repeats, R: reverse repeats, C: complementary repeats. (E) The number of oligonucleotide repeats in different regions. LSC: large single-copy region, SSC: small single-copy region, IR: inverted repeat region, LSC/IR: repeat sequences crossed LSC and IR regions, etc. (F) The number of repeats in different repeat units.

The oligonucleotide repeat sequences were also analyzed using the REPuter program to detect the abundance of four types of oligonucleotide repeat, including forward (F), palindromic (P), reverse (R), and complementary (C). Although minor variations presented about the total number of oligonucleotide repeats, the distribution of the four types of repeats and the size of the repeats presented an obvious resemblance. In terms of the number of oligonucleotide repeats and their distribution in each type, we found 42 repeats (F = 16, P = 22, R = 4) in the cp genome of *B. hispida*; 41 (F = 16, P = 21, R = 4) in *C. lanatus*; 46 (F = 14, P = 26, R = 4) in *L. siceraria*; and 42 (F = 14, P = 23, R = 5) in *C. colocynthis* (Figure 2A). The length of repeats was mostly found between 20 and 24 bp (Figure 2C). The palindromic repeats were the most abundant repeats, followed by forward repeats, whereas the number of reverse repeats was low. None of the species had complementary repeats. We also located the region of each oligonucleotide repeat; significant consistency was presented among the four species. The number was exactly the same in all the species regarding the repeats located in the IR regions (6) and some shared sequences, including sequences between LSC and IRa/b (4), between SSC and IRa/b (2), and from IRb to IRa crossing SSC (Figure 2B).

3.5. IR Contraction and Expansion

The genome length of the chloroplast ranged from 159,758 bp (*B. hispida*) to 157,147 bp (*C. colocynthis*). Furthermore, in the cp genome of *B. hispida*, the length of the IR regions was the shortest with 260,080 bp, while that of the SSC region was the longest with 180,060 bp (Table S5). Thus, we inferred that the variation in size of the cp genome was contributed to by the expansion and contraction of the IR regions (Figure 3). The junction sites between each region were denoted as: J_{LB} (IRb/LSC), J_{SA} (SSC/IRa), J_{SB} (IRb/SSC), and J_{LA} (IRa/LSC). All eight species analyzed presented functional *ycf1* genes, six of which were at J_{SA}, while the other two were located in the SSC region completely. Moreover, the *ycf1*Ψ (pseudo-copy) was only present in two species (*B. hispida* and *L. siceraria*) at J_{SB} and were 3 bp and 25 bp in the SSC region, respectively. The *ndhF* gene was revealed in all species in J_{SB} with the same length (2246 bp) and relatively consistent position with only a few bp in the IRb region, except *C. hystrix*, with 21 bp. and *B. hispida* (completely located in the SSC region).

The *rpl2* gene was found close to J_{LB}, while that of two species (*C. moschata* and *C. lanatus*) were in the LSC region with 11 bp and 6 bp, respectively. At the same time, the duplicate *rpl2* genes were absent in the same two specific species. The *rps19* gene was the most variant gene among all the genes close to the IR junction. In the four species, the *rps19* genes were 2 bp in the IRb region and the remaining four were completely in LSC region.

3.6. Divergence Analysis of Chloroplast Genome

To identify the diversity in the chloroplast genomes of four Benincaseae species, we visualized the percentage of identity between the sequences and colored regions of high conservation using mVISTA program. As shown in Figure 4, the sequences varied remarkably among different regions. Firstly, most of the differences were located in the LSC and SSC regions, while the IR regions were almost identical among the four species except the *rps12* gene, revealing that the IR regions were more conservative than the LSC and SSC regions. Moreover, the IGS regions revealed themselves to be remarkably more divergent than the gene regions. Notable divergent non-coding regions included: *trnR-UCU-atpA*, *atpH-atpI*, *trnT-GGU-psbD*, *trnL-UAA-trnF-GAA*, *accD-pasI*, and *ndhF-rpl32*. Genes such as *ycf1*, *ycf2*, *accD*, *psbA*, *ccsA*, *ndhF*, and *matK* were found to be highly divergent coding genes.

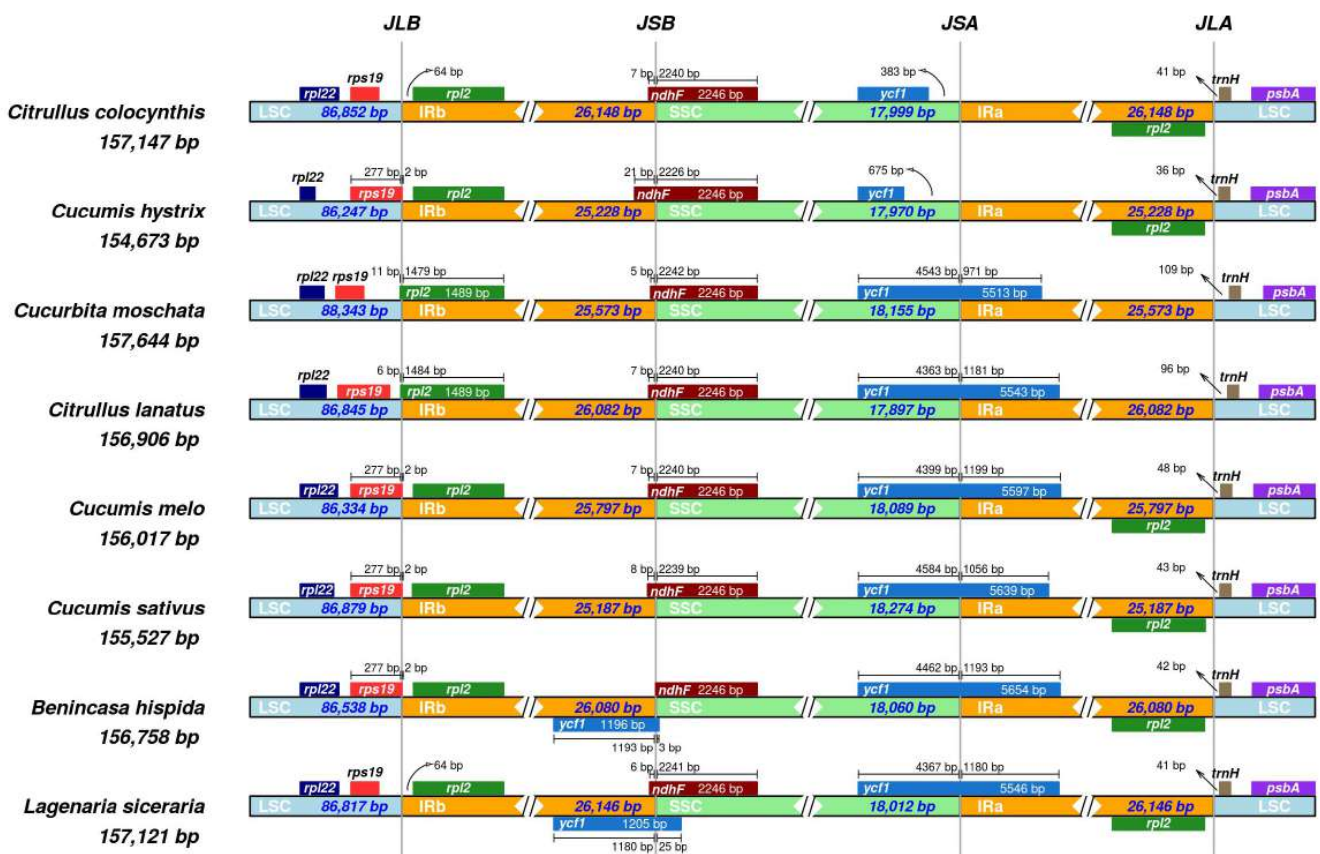


Figure 3. Comparison of junctions between the LSC, SSC, and IRs among eight species. Number above indicates the distance in bp between the ends of the genes and the border sites (distances are not to scale in this figure).

The K_a/K_s ratio is an essential index to identify a mutation as neutral, purifying, or beneficial. Thus, we compared *B. hispida* with *C. colocynthis*, *C. lanatus*, and *L. siceraria*, respectively, to analyze the synonymous substitutions (K_s), the non-synonymous substitutions (K_a), and their ratio (K_a/K_s) of 73 PCGs (Table S6). In total, 18 genes could not be determined due to absent information ($K_s = 0$). After deleting these genes, as well as the non-substitution results, we found that the genes carrying out photosynthesis functions revealed $K_a/K_s = 0$ or at relatively low values, indicating that these groups of genes were fairly conserved. The K_a/K_s ratio of 26 genes was lower than 0.5 and that of 96% genes was lower than 1, with only 5 exceptions (*accD*, *clpP*, *rps4*, *ycf1*, and *ycf2*). We then performed a purifying/positive selection site evaluation for these five genes (Table S7). However, only two genes, *accD* and *rps4*, presented sites potentially under positive selection, indicated by the high empirical Bayes values (Table 3).

To obtain a holistic understanding of the sequence divergence, we performed a sliding window analysis to visualize the nucleotide variability values of all the cp genomes. We found that none of the π values of the CDS genes exceeded 0.05 and that the IGSs were more divergent than the gene regions, which was consistent with the aforementioned analysis. It can be clearly seen in the figure that the SSC and LSC regions were much more divergent than the IR regions, the π value of which was remarkably low and mirror-symmetrized with SSC as the center (Figure 5).

Table 3. Likelihood ratio tests of five potential genes under positive selection.

Gene Name	Models	np	ln L	Likelihood Ratio Test <i>p</i> -Value	Positively Selected Sites	
					AA-Site	Score
<i>accD</i>	M8 (beta)	10	−2173.400149	0.007931755	159 W	0.984 *
	M7 (beta & $\omega > 1$)	8	−2178.23703			
<i>clpP</i>	M8 (beta)	10	−926.578492	0.070969008		
	M7 (beta & $\omega > 1$)	8	−929.224004			
<i>rps4</i>	M8 (beta)	10	−877.349259	0.030217199	158 Q	0.971 *
	M7 (beta & $\omega > 1$)	8	−880.848603			
<i>ycf1</i>	M8 (beta)	10	−6139.981658	0.156641895		
	M7 (beta & $\omega > 1$)	8	−6141.835451			
<i>ycf2</i>	M8 (beta)	10	−9230.970637	0.063743376		
	M7 (beta & $\omega > 1$)	8	−9233.723527			

np represents degree of freedom; ln L represents log likelihood values; *: empirical Bayes values > 0.95.

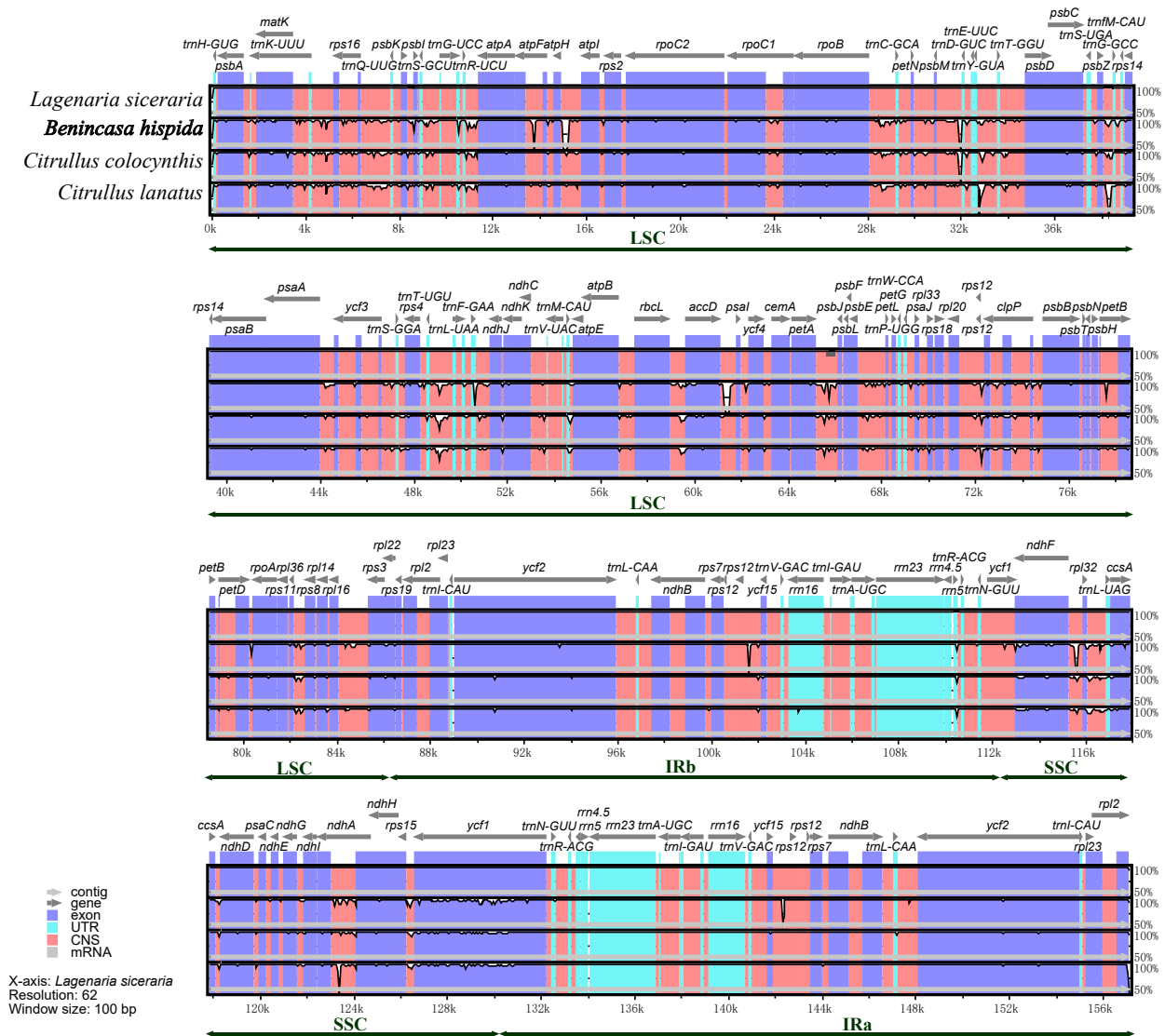


Figure 4. Sequence identity plot comparing the chloroplast genomes among *Benincaseae* species with *Lagenaria siceraria* set as a reference using mVISTA. Pink bars represent noncoding sequences (CNS), and white peaks represent genome divergence. The y-axis represents the percentage identity (shown: 50–100%).

3.7. Phylogenetic Analysis

To locate the phylogenetic position of *B. hispida* precisely, we selected 26 species (Table S8) and constructed two phylogenetic trees using the complete cp genome (Figure 6A) and 73 selected CDSs (Figure 6B), respectively. The results all suggested that *B. hispida* was closely related with *Cucumis*, *Citrullus*, and *Lagenaria* as their sister group, with fairly high bootstrap values. The phylogenetic relationship results of the two approaches presented high consistency, with two main variations. Firstly, in general, the bootstrap values in the tree that applied the complete cp genome were higher than in the tree constructed with 73 CDSs (Figure 6B). In addition, Begoniaceae was a sister group with Coriariaceae and Corynocarpaceae, according to Figure 6A, while in Figure 6B, Coriariaceae and Corynocarpaceae were the early-diverging lineages of Begoniaceae. However, only 82 bootstrap values supported the former situation (Figure 6A) while 94 supported the second (Figure 6B).

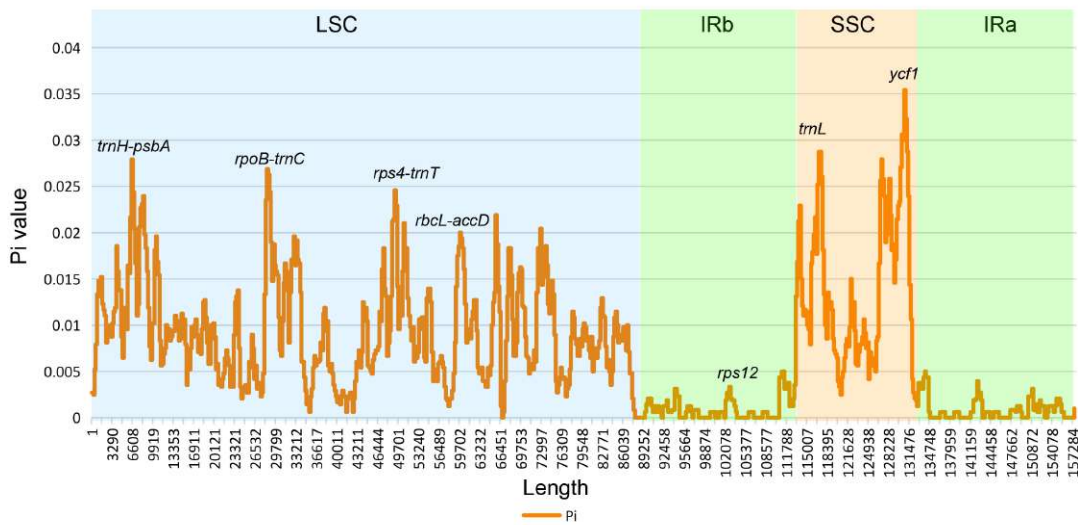


Figure 5. Nucleotide diversity (π) values among the *Benincaseae* species.

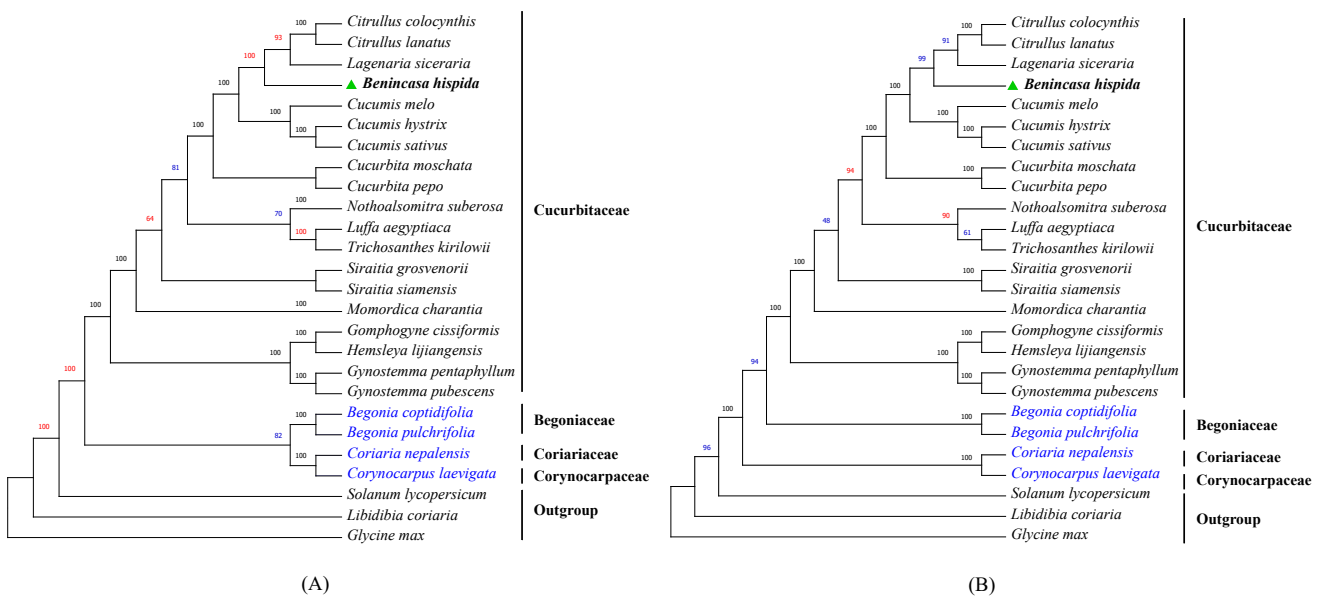


Figure 6. Maximum likelihood (ML) tree of Cucurbitales. (A) The phylogenetic tree constructed by complete chloroplast genome of 23 species. (B) The phylogenetic tree builds with 72 genes. The positions of *Benincasa hispida* are marked with green triangles. Numbers above branches are bootstrap values, and the bootstrap values higher or lower than those of the other tree are marked as red or blue, respectively. *Glycine max* set as the root in both trees.

4. Discussion

In this study, we sequenced and reported the complete chloroplast genome of *Benincasa hispida* and performed comparative analyses with three other, closely related species selected from the *Benincaseae* but distinct enough to obtain reasonable results, providing valuable genetic data for phylogeographic and population genetic investigation [50,51].

The cp genome revealed high consistency in terms of its quadruple structure, gene content, and organization not only in *Benincaseae* [52,53], but also in other angiosperms [54]. The genome size differed less than 400 bp, with almost identical gene numbers, signifying that the cp genomes among the four analyzed species were conservative on the whole. The GC content of *B. hispida* varied across different regions and functions. The rRNA sequences were considerably rich in GC bases; as a consequence, the IR regions rich in rRNA appear to have had higher GC content than the other regions. These findings agree with those of previous studies [55,56].

However, changes were found that provide valuable information for understanding the development and evolution [50,57]. The bias of the codon usage in the plant cp genome was an important evolutionary feature for the studies regarding mRNA translation, new gene discovery, and molecular biology [58]. Previous studies have confirmed that genes tend to choose preferred synonymous codons for specific amino acids rather than randomly distributions [59,60]. Our study showed that genes of *B. hispida* prefer codons with A/T in the third position, which was consistent with previous studies [61,62].

Microsatellites, or SSRs, are widely distributed in cp genomes that serve as molecular markers for phylogenetic relationship inference [63,64]. Moreover, SSRs are also related to different types of genome rearrangement, recombinations, and large inversions [65,66]. Similar to previous studies, we found that mononucleotide repeats were the most abundant types of repeat and that their numbers in the LSC region far surpassed those in the SSC and IR regions [67]. Furthermore, a greater number of palindromic repeats were found among four types of repeat, while previous studies revealed that the forward repeats were the most abundant repeats [61,68]. We specifically analyzed the abundance of SSRs that differed from gene regions to intronic gene regions and verified that the IGSs contained much higher SSR density than the others. Thus, we inferred that IGS regions may undergo gene rearrangement and recombination more frequently than gene regions. Moreover, our results support the hypothesis that cpSSRs are more often composed by polyA or polyT than polyG or polyC [69,70], implying that IGSs might be relatively rapidly mutating regions [71,72].

It is commonly agreed that variations in genome size in the chloroplast are the consequence of IR contraction and expansion, leading to gene duplication and deletion and the presence of pseudogenes [68,73]. We found that *ycf1*Ψ pseudogenes were only detected in *B. hispida* and *L. siceraria*, which were also sister groups in the ML phylogenetic trees. Furthermore, no *rps19*Ψ pseudogenes were observed in any of the species analyzed; their presence was thought to be responsible for the loss of function of the *rps19* gene [74,75]. These results imply that gene variation at IR boundaries may contribute to the understanding of the cp genome at a molecular level and serve as an indicator for evolutionary investigation [76,77].

It is worthwhile to study the genetic diversity among the four *Benincaseae* species because the chloroplast genome plays a crucial role in the study of phylogeny, gene flow between species, and population genetics among different species. [64,78]. The coding regions were generally found to be more conserved than the non-coding regions. Furthermore, some of the coding genes, namely the *ycf1*, *ycf2*, *matK*, *accD*, and *ndhF* genes, were commonly found to be relatively divergent [79,80]. In addition, the LSC and SSC regions were further confirmed to be more divergent in comparison with the IR regions [81]. We also discovered that genes related to photosynthesis with low *Ka/Ks* ratios showed slow evolution rates, while functional genes, such as *accD*, revealed high evolutionary rates, indicating that genes carrying out vital functions were conserved and vice versa [74,82].

Among the five genes that showed Ka/Ks values greater than one, two genes, *accD* and *rps4*, presented one positive selection site, respectively. These results indicate that the *accD* gene may have changed under evolutionary pressure [83]

Currently, protein-coding genes are commonly implemented for phylogenetic tree building [84]. While the complete cp genome contains richer information but requires a longer time to perform, higher-end equipment and the population distance may be exaggerated for the highly divergent features of IGS genes [85]. In this study, we applied both methods to build the phylogenetic trees. The first tree, built with the complete cp genome, revealed higher bootstrap values in general, while the other tree, built with coding genes, showed a slightly different phylogenetic order in four species, out of twenty-six in total. In general, the phylogenetic position revealed was consistent with previous studies [86–89]. However, the phylogenetic relationships we discovered within Cucurbitaceae differed from the results of previous phylogenetic marker-based taxonomy research [90,91]. This may have been due to the different approaches used for phylogenetic tree construction, with further investigation needed.

In conclusion, our study first shed light on the structure and content of the cp genome of *B. hispida*, an economically important fruit crop widely distributed in several tropical countries and extensively consumed worldwide. We also offered information regarding similarities and divergence, enriching the understanding of the species of Benincaseae. Moreover, information about highly polymorphic regions was also provided regarding molecular markers and highly divergent regions, which might be useful for further studies of the taxonomy and phylogeographics of Benincaseae subfamilies.

Supplementary Materials: The following are available online at <https://www.mdpi.com/article/10.3390/genes13030461/s1>. Table S1: Intron details. Table S2: codon usage details. Table S3: RNA editing sites raw data. Table S4: Detail of SSR and long repeats. Table S5: Sequence information of four species. Table S6: Ka/Ks raw data. Table S7: Evolutionary tree species information. Table S8: Information of 26 species sequences.

Author Contributions: Supervision, C.S. and S.W.; validation, W.S. and S.W.; visualization, Z.C. and L.H.; writing—original draft, W.S. and Z.C.; writing—review and editing, W.S.; data curation, H.Z. and G.D.; funding acquisition, L.H. and C.S.; investigation, G.D.; methodology, W.S. and Z.C.; formal analysis, Q.F.; project administration, W.S. and C.S.; resources, G.D. and L.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (NO. 31801022 and NO. 31701090) and Shandong Province Natural Science Foundation of China (NO. ZR2019BC094).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are openly available in the GenBank of NCBI at <https://www.ncbi.nlm.nih.gov> (accessed on 13 January 2022), reference number (MW362306).

Conflicts of Interest: The authors declare that they have no conflict of interest.

References

1. Christenhusz, M.J.M.; Byng, J.W. The number of known plants species in the world and its annual increase. *Phytotaxa* **2016**, *261*, 201–217. [CrossRef]
2. Ward, B.L.; Anderson, R.S.; Bendich, A.J. The mitochondrial genome is large and variable in a family of plants (*Cucurbitaceae*). *Cell* **1981**, *25*, 793–803. [CrossRef]
3. Nandecha, C.; Nahata, A.; Dixit, V.K. Effect of *Benincasa hispida* fruits on testosterone-induced prostatic hypertrophy in albino rats. *Curr. Ther. Res.* **2010**, *71*, 331–343. [CrossRef]
4. Kocyan, A.; Zhang, L.-B.; Schaefer, H.; Renner, S.S. A multi-locus chloroplast phylogeny for the *Cucurbitaceae* and its implications for character evolution and classification. *Mol. Phylogenet. Evol.* **2007**, *44*, 553–577. [CrossRef]

5. Renner, S.S.; Schaefer, H. Phylogeny and Evolution of the *Cucurbitaceae*. In *Genetics and Genomics of Cucurbitaceae. Plant Genetics and Genomics: Crops and Models*; Grumet, R., Katzir, N., Garcia-Mas, J., Eds.; Springer: Cham, The Netherlands, 2016; Volume 20, pp. 155–172. [[CrossRef](#)]
6. Guo, J.; Xu, W.; Hu, Y.; Huang, J.; Zhao, Y.; Zhang, L.; Huang, C.-H.; Ma, H. Phylotranscriptomics in *Cucurbitaceae* reveal multiple whole-genome duplications and key morphological and molecular innovations. *Mol. Plant* **2020**, *13*, 1117–1133. [[CrossRef](#)]
7. Steward, F.C. Some Economic Plants: Tropical Crops: Dicotyledons. J. W. Purseglove. Wiley, New York, 1968; 2 vols., xx + 719 pp., illus. \$8.50 each. *Science* **1969**, *163*, 1050–1051. [[CrossRef](#)]
8. Thomas, T.D.; Sreejesh, K.R. Callus induction and plant regeneration from cotyledonary explants of ash gourd (*Benincasa hispida* L.). *Sci. Hortic.* **2004**, *100*, 359–367. [[CrossRef](#)]
9. Naik, R.; Buha, M.; Acharya, R.; Borkar, S.D. Role of vegetables (*Shaka Dravyas*) in prevention and management of gastro—Intestinal tract diseases: A critical review. *J. Res. Tradit. Med.* **2016**, *2*, 103–112. [[CrossRef](#)]
10. Al-Snafi, A.E. The Pharmacological Importance of *Benincasa hispida*. A review. *Int. J. Pharma Sci. Res.* **2013**, *4*, 975–9492. Available online: https://www.researchgate.net/publication/313676687_The_Pharmacological_Importance_of_Benincasa_hispida_A_review (accessed on 13 December 2021).
11. Rachchh, M.A.; Jain, S.M. Gastroprotective effect of *Benincasa hispida* fruit extract. *Indian J. Pharmacol.* **2008**, *40*, 271–275. [[CrossRef](#)] [[PubMed](#)]
12. Dhingra, D.; Joshi, P. Antidepressant-like activity of *Benincasa hispida* fruits in mice: Possible involvement of monoaminergic and GABAergic systems. *J. Pharmacol. Pharmacother.* **2012**, *3*, 60–62. [[CrossRef](#)]
13. Jayasree, T.; Kishore, K.K.; Vinay, M.; Vasavi, P.; Dixit, R. Evaluation of the diuretic effect of the chloroform extract of the *Benincasa hispida* rind (pericarp) extract in guinea-pigs. *J. Clin. Diagn. Res.* **2011**, *5*, 578–582.
14. Lee, K.-H.; Choi, H.-R.; Kim, C.-H. Anti-angiogenic effect of the seed extract of *Benincasa hispida* Cogniaux. *J. Ethnopharmacol.* **2005**, *97*, 509–513. [[CrossRef](#)] [[PubMed](#)]
15. Qadrie, Z.L.; Hawisa, N.T.; Khan, M.W.A.; Samuel, M.; Anandan, R. Antinociceptive and anti-pyretic activity of *Benincasa hispida* (thunb.) cogn. In Wistar albino rats. *Pak. J. Pharm. Sci.* **2009**, *22*, 287–290. [[CrossRef](#)] [[PubMed](#)]
16. 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](#)] [[PubMed](#)]
17. Natarajan, D.; Lavarasan, R.J.; Babu, S.C.; Refai, M.; Ansari, L. Antimicrobial studies on methanol extract of *Benincasa hispida* cogn., fruit. *Anc. Sci. Life* **2003**, *22*, 98–100. [[PubMed](#)]
18. Bimakar, M.; Rahman, R.A.; Taip, F.S.; Adzahan, N.M.; Sarker, M.Z.I.; Ganjloo, A. Optimization of ultrasound-assisted extraction of crude oil from winter melon (*Benincasa hispida*) seed using response surface methodology and evaluation of its antioxidant activity, total phenolic content and fatty acid composition. *Molecules* **2012**, *17*, 11748–11762. [[CrossRef](#)] [[PubMed](#)]
19. Grover, J.K.; Adiga, G.; Vats, V.; Rathi, S.S. Extracts of *Benincasa hispida* prevent development of experimental ulcers. *J. Ethnopharmacol.* **2001**, *78*, 159–164. [[CrossRef](#)]
20. Palmer, J.D. Plastid chromosomes: Structure and evolution. *Mol. Biol. Plast.* **1991**, *7*, 5–53. [[CrossRef](#)]
21. Ahmed, I.; Biggs, P.J.; Matthews, P.J.; Collins, L.J.; Hendy, M.D.; Lockhart, P.J. Mutational dynamics of aroid chloroplast genomes. *Genome Biol. Evol.* **2012**, *4*, 1316–1323. [[CrossRef](#)]
22. Sloan, D.B.; Triant, D.A.; Forrester, N.J.; Bergner, L.M.; Wu, M.; Taylor, D.R. A recurring syndrome of accelerated plastid genome evolution in the angiosperm tribe *Sileneae* (*Caryophyllaceae*). *Mol. Phylogenetics Evol.* **2013**, *72*, 82–89. [[CrossRef](#)]
23. Ahmed, I. Chloroplast genome sequencing: Some reflections. *J. Next Gener. Seq. Appl.* **2015**, *2*, 2469–9853. [[CrossRef](#)]
24. Lössl, A.G.; Waheed, M.T. Chloroplast-derived vaccines against human diseases: Achievements, challenges and scopes. *Plant Biotechnol. J.* **2011**, *9*, 527–539. [[CrossRef](#)] [[PubMed](#)]
25. Xie, D.; Xu, Y.; Wang, J.; Liu, W.; Zhou, Q.; Luo, S.; Huang, W.; He, X.; Li, Q.; Peng, Q.; et al. The wax gourd genomes offer insights into the genetic diversity and ancestral cucurbit karyotype. *Nat. Commun.* **2019**, *10*, 5158. [[CrossRef](#)] [[PubMed](#)]
26. Doyle, J.J.; Doyle, J.L. Isolation of plant DNA from fresh tissue. *Focus* **1990**, *12*, 13–15.
27. Wang, Y.; Wang, S.; Liu, Y.; Yuan, Q.; Sun, J.; Guo, L. Chloroplast genome variation and phylogenetic relationships of *Atractylodes* species. *BMC Genom.* **2021**, *22*, 103. [[CrossRef](#)]
28. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [[CrossRef](#)]
29. Andrews, S. FastQC: A quality control tool for high throughput sequence data. In *Babraham Bioinformatics*; Babraham Institute: Cambridge, UK, 2010; Available online: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (accessed on 13 December 2021).
30. Ewels, P.; Magnusson, M.; Lundin, S.; Käller, M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **2016**, *32*, 3047–3048. [[CrossRef](#)] [[PubMed](#)]
31. Guo, L.; Guo, S.; Xu, J.; He, L.; Carlson, J.E.; Hou, X. Phylogenetic analysis based on chloroplast genome uncover evolutionary relationship of all the nine species and six cultivars of tree peony. *Ind. Crops Prod.* **2020**, *153*, 112567. [[CrossRef](#)]
32. Luo, R.; Liu, B.; Xie, Y.; Li, Z.; Huang, W.; Yuan, J.; He, G.; Chen, Y.; Pan, Q.; Liu, Y.; et al. SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *GigaScience* **2012**, *1*, 18. [[CrossRef](#)] [[PubMed](#)]
33. Muraguri, S.; Xu, W.; Chapman, M.; Muchugi, A.; Oluwaniyi, A.; Oyeboji, O.; Liu, A. Intraspecific variation within *Castor bean* (*Ricinus communis* L.) based on chloroplast genomes. *Ind. Crop. Prod.* **2020**, *155*, 112779. [[CrossRef](#)]

34. Dierckxsens, N.; Mardulyn, P.; Smits, G. NOVOPlasty: De novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* **2017**, *45*, e18. [CrossRef] [PubMed]
35. 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]
36. Lowe, T.M.; Eddy, S.R. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **1997**, *25*, 0955–964. [CrossRef]
37. Lagesen, K.; Hallin, P.; Rødland, E.A.; Staerfeldt, H.-H.; Rognes, T.; Ussery, D.W. RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **2007**, *35*, 3100–3108. [CrossRef]
38. Lehwark, P.; Greiner, S. GB2sequin—A file converter preparing custom GenBank files for database submission. *Genomics* **2019**, *111*, 759–761. [CrossRef]
39. Lohse, M.; Drechsel, O.; Bock, R. OrganellarGenomeDRAW (OGDRAW): A tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Curr. Genet.* **2007**, *52*, 267–274. [CrossRef]
40. Kearse, M.; Moir, R.; Wilson, A.; Stones-Havas, S.; Cheung, M.; Sturrock, S.; Buxton, S.; Cooper, A.; Markowitz, S.; Duran, C.; et al. Geneious Basic: Bioinformatics Software for Sequence Data Analysis. 2012. Available online: <https://www.geneious.com/> (accessed on 13 December 2021).
41. Kumar, S.; Stecher, G.; Li, M.; Nnyaz, C.; Tamura, K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **2018**, *35*, 1547–1549. [CrossRef] [PubMed]
42. Mower, J.P. The PREP suite: Predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments. *Nucleic Acids Res.* **2009**, *37*, W253–W259. [CrossRef]
43. Beier, S.; Thiel, T.; Münch, T.; Scholz, U.; Mascher, M. MISA-web: A web server for microsatellite prediction. *Bioinformatics* **2017**, *33*, 2583–2585. [CrossRef] [PubMed]
44. Kurtz, S.; Choudhuri, J.V.; Ohlebusch, E.; Schleiermacher, C.; Stoye, J.; Giegerich, R. REPuter: The manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* **2001**, *29*, 4633–4642. [CrossRef] [PubMed]
45. Amiryousefi, A.; Hyvönen, J.; Poczai, P. Irscope: An online program to visualize the junction sites of chloroplast genomes. *Bioinformatics* **2018**, *34*, 3030–3031. [CrossRef]
46. Rozas, J.; Ferrer-Mata, A.; Sánchez-DelBarrio, J.C.; Guirao-Rico, S.; Librado, P.; Ramos-Onsins, S.E.; Sánchez-Gracia, A. DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol. Biol. Evol.* **2017**, *34*, 3299–3302. [CrossRef]
47. Yang, Z.; Nielsen, R.; Goldman, N.; Pedersen, A.-M.K. Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites. *Genetics* **2000**, *155*, 431–449. [CrossRef]
48. Gao, F.; Chen, C.; Arab, D.A.; Du, Z.; He, Y.; Ho, S.Y.W. EasyCodeML: A visual tool for analysis of selection using CodeML. *Ecol. Evol.* **2019**, *9*, 3891–3898. [CrossRef]
49. Zhu, J.; Wen, D.; Yu, Y.; Meudt, H.M.; Nakhleh, L. Bayesian inference of phylogenetic networks from bi-allelic genetic markers. *PLOS Comput. Biol.* **2018**, *14*, e1005932. [CrossRef] [PubMed]
50. Shaw, J.; Lickey, E.B.; Schilling, E.E.; Small, R.L. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: The tortoise and the hare III. *Am. J. Bot.* **2007**, *94*, 275–288. [CrossRef] [PubMed]
51. Poczai, P.; Hyvönen, J. The complete chloroplast genome sequence of the CAM epiphyte Spanish moss (*Tillandsia usneoides*, Bromeliaceae) and its comparative analysis. *PLoS ONE* **2017**, *12*, e0187199. [CrossRef] [PubMed]
52. Hu, J.-B.; Zhou, X.-Y.; Li, J.-W. Development of novel chloroplast microsatellite markers for Cucumis from sequence database. *Biol. Plant.* **2009**, *53*, 793–796. [CrossRef]
53. Bhowmick, B.K.; Jha, S. Differential heterochromatin distribution, flow cytometric genome size and meiotic behavior of chromosomes in three *Cucurbitaceae* species. *Sci. Hort.* **2015**, *193*, 322–329. [CrossRef]
54. Bausher, M.G.; Singh, N.D.; Lee, S.-B.; Jansen, R.K.; Daniell, H. The complete chloroplast genome sequence of *Citrus sinensis* (L.) Osbeck var ‘Ridge Pineapple’: Organization and phylogenetic relationships to other angiosperms. *BMC Plant Biol.* **2006**, *6*, 21. [CrossRef] [PubMed]
55. Mehmood, F.; Abdullah, S.I.; Ahmed, I.; Waheed, M.T.; Mirza, B. Characterization of *Withania somnifera* chloroplast genome and its comparison with other selected species of Solanaceae. *Genomics* **2020**, *112*, 1522–1530. [CrossRef] [PubMed]
56. Guo, S.; Guo, L.; Zhao, W.; Xu, J.; Li, Y.; Zhang, X.; Shen, X.; Wu, M.; Hou, X. Complete chloroplast genome sequence and phylogenetic analysis of *Paeonia ostii*. *Molecules* **2018**, *23*, 246. [CrossRef]
57. Daniell, H.; Jin, S.; Zhu, X.; Gitzendanner, M.A.; Soltis, D.E.; Soltis, P.S. Green giant—A tiny chloroplast genome with mighty power to produce high-value proteins: History and phylogeny. *Plant Biotechnol. J.* **2021**, *19*, 430–447. [CrossRef]
58. Yang, X.; Luo, X.; Cai, X. Analysis of codon usage pattern in *Taenia saginata* based on a transcriptome dataset. *Parasites Vectors* **2014**, *7*, 527. [CrossRef] [PubMed]
59. Sorimachi, K. Codon evolution in double-stranded organelle DNA: Strong regulation of homonucleotides and their analog alternations. *Nat. Sci.* **2010**, *2*, 846–854. [CrossRef]
60. Li, W.; Zhang, C.; Guo, X.; Liu, Q.; Wang, K. Complete chloroplast genome of *Camellia japonica* genome structures, comparative and phylogenetic analysis. *PLoS ONE* **2019**, *14*, e0216645. [CrossRef] [PubMed]
61. Saina, J.K.; Li, Z.-Z.; Gichira, A.W.; Liao, Y.-Y. The complete chloroplast genome sequence of tree of heaven (*Ailanthus altissima* (Mill.) (Sapindales: Simaroubaceae), an important pantropical tree. *Int. J. Mol. Sci.* **2018**, *19*, 929. [CrossRef] [PubMed]

62. Wang, W.; Yu, H.; Wang, J.; Lei, W.; Gao, J.; Qiu, X.; Wang, J. The complete chloroplast genome sequences of the medicinal plant *Forsythia suspensa* (Oleaceae). *Int. J. Mol. Sci.* **2017**, *18*, 2288. [[CrossRef](#)]
63. Ahmed, I.; Matthews, P.J.; Biggs, P.; Naeem, M.; McLenachan, P.A.; Lockhart, P.J. Identification of chloroplast genome loci suitable for high-resolution phylogeographic studies of *Colocasia esculenta* (L.) Schott (Araceae) and closely related taxa. *Mol. Ecol. Resour.* **2013**, *13*, 929–937. [[CrossRef](#)] [[PubMed](#)]
64. Cavender-Bares, J.; González-Rodríguez, A.; Eaton, D.A.R.; Hipp, A.A.L.; Beulke, A.; Manos, P.S. Phylogeny and biogeography of the American live oaks (*Quercus* subsection *Virentes*): A genomic and population genetics approach. *Mol. Ecol.* **2015**, *24*, 3668–3687. [[CrossRef](#)]
65. Guisinger, M.M.; Kuehl, J.V.; Boore, J.L.; Jansen, R.K. Extreme reconfiguration of plastid genomes in the angiosperm family geraniaceae: Rearrangements, repeats, and codon usage. *Mol. Biol. Evol.* **2010**, *28*, 583–600. [[CrossRef](#)]
66. Song, Y.; Zhang, Y.; Xu, J.; Li, W.; Li, M. Characterization of the complete chloroplast genome sequence of *Dalbergia* species and its phylogenetic implications. *Sci. Rep.* **2019**, *9*, 20401. [[CrossRef](#)]
67. Jeon, J.-H.; Kim, S.-C. Comparative analysis of the complete chloroplast genome sequences of three closely related east-Asian wild roses (*Rosa* sect. *Synstylae*; Rosaceae). *Genes* **2019**, *10*, 23. [[CrossRef](#)] [[PubMed](#)]
68. Abdullah, S.I.; Mehmood, F.; Ali, Z.; Malik, M.S.; Waseem, S.; Mirza, B.; Ahmed, I.; Waheed, M.T. Comparative analyses of chloroplast genomes among three Firmiana species: Identification of mutational hotspots and phylogenetic relationship with other species of *Malvaceae*. *Plant Gene* **2019**, *19*, 100199. [[CrossRef](#)]
69. Shen, X.; Wu, M.; Liao, B.; Liu, Z.; Bai, R.; Xiao, S.; Li, X.; Zhang, B.; Xu, J.; Chen, S. Complete chloroplast genome sequence and phylogenetic analysis of the medicinal plant *Artemisia annua*. *Molecules* **2017**, *22*, 1330. [[CrossRef](#)]
70. Raubeson, L.A.; Peery, R.; Chumley, T.W.; Dziubek, C.; Fourcade, H.M.; Boore, J.L.; Jansen, R.K. Comparative chloroplast genomics: Analyses including new sequences from the angiosperms *Nuphar advena* and *Ranunculus macranthus*. *BMC Genom.* **2007**, *8*, 174. [[CrossRef](#)]
71. Provan, J.; Powell, W.; Hollingsworth, P.M. Chloroplast microsatellites: New tools for studies in plant ecology and evolution. *Trends Ecol. Evol.* **2001**, *16*, 142–147. [[CrossRef](#)]
72. Liu, L.; Wang, Y.; He, P.; Li, P.; Lee, J.; Soltis, D.E.; Fu, C. Chloroplast genome analyses and genomic resource development for epilithic sister genera *Oresitrophe* and *Mukdenia* (Saxifragaceae), using genome skimming data. *BMC Genom.* **2018**, *19*, 235. [[CrossRef](#)]
73. Zhu, B.; Qian, F.; Hou, Y.; Yang, W.; Cai, M.; Wu, X. Complete chloroplast genome features and phylogenetic analysis of *Eruca sativa* (Brassicaceae). *PLoS ONE* **2021**, *16*, e0248556. [[CrossRef](#)]
74. Menezes, A.P.A.; Resende-Moreira, L.C.; Buzatti, R.S.O.; Nazareno, A.G.; Carlsen, M.; Lobo, F.P.; Kalapothakis, E.; Lovato, M.B. Chloroplast genomes of *Byrsonima* species (*Malpighiaceae*): Comparative analysis and screening of high divergence sequences. *Sci. Rep.* **2018**, *8*, 2210. [[CrossRef](#)] [[PubMed](#)]
75. Shahzadi, I.; Abdullah, M.F.; Ali, Z.; Ahmed, I.; Mirza, B. Chloroplast genome sequences of *Artemisia maritima* and *Artemisia absinthium*: Comparative analyses, mutational hotspots in genus *Artemisia* and phylogeny in family Asteraceae. *Genomics* **2020**, *112*, 1454–1463. [[CrossRef](#)]
76. Nazareno, A.G.; Carlsen, M.; Lohmann, L.G. Complete chloroplast genome of *Tanaecium tetragonolobum*: The first ignoniaceae plastome. *PLoS ONE* **2015**, *10*, e0129930. [[CrossRef](#)] [[PubMed](#)]
77. Jansen, R.K.; Sasaki, 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 rpl22 to the nucleus. *Mol. Biol. Evol.* **2011**, *28*, 835–847. [[CrossRef](#)] [[PubMed](#)]
78. Li, X.; Li, Y.; Zang, M.; Li, M.; Fang, Y. Complete chloroplast genome sequence and phylogenetic analysis of *Quercus acutissima*. *Int. J. Mol. Sci.* **2018**, *19*, 2443. [[CrossRef](#)]
79. Amiryousefi, A.; Hyvönen, J.; Poczai, P. The chloroplast genome sequence of bittersweet (*Solanum dulcamara*): Plastid genome structure evolution in Solanaceae. *PLoS ONE* **2018**, *13*, e0196069. [[CrossRef](#)]
80. Du, Y.-P.; Bi, Y.; Yang, F.-P.; Zhang, M.-F.; Chen, X.-Q.; Xue, J.; Zhang, X.-H. Complete chloroplast genome sequences of *Lilium*: Insights into evolutionary dynamics and phylogenetic analyses. *Sci. Rep.* **2017**, *7*, 5751. [[CrossRef](#)] [[PubMed](#)]
81. Huo, Y.; Gao, L.; Liu, B.; Yang, Y.; Kong, S.; Sun, Y.; Yang, Y.; Wu, X. Complete chloroplast genome sequences of four *Allium* species: Comparative and phylogenetic analyses. *Sci. Rep.* **2019**, *9*, 12250. [[CrossRef](#)]
82. Xiao-Ming, Z.; Junrui, W.; Li, F.; Sha, L.; Hongbo, P.; Lan, Q.; Jing, L.; Yan, S.; Weihua, Q.; Lifang, Z.; et al. Inferring the evolutionary mechanism of the chloroplast genome size by comparing whole-chloroplast genome sequences in seed plants. *Sci. Rep.* **2017**, *7*, 1555. [[CrossRef](#)]
83. Kode, V.; Mudd, E.A.; Iamtham, S.; Day, A. The tobacco plastid accD gene is essential and is required for leaf development. *Plant J.* **2005**, *44*, 237–244. [[CrossRef](#)] [[PubMed](#)]
84. Cui, Y.; Nie, L.; Sun, W.; Xu, Z.; Wang, Y.; Yu, J.; Song, J.; Yao, H. Comparative and phylogenetic analyses of ginger (*Zingiber officinale*) in the family *Zingiberaceae* based on the complete chloroplast genome. *Plants* **2019**, *8*, 283. [[CrossRef](#)]
85. Cheng, Y.; Zhang, L.; Qi, J.; Zhang, L. Complete chloroplast genome sequence of *Hibiscus cannabinus* and comparative analysis of the *Malvaceae* family. *Front. Genet.* **2020**, *11*, 277. [[CrossRef](#)] [[PubMed](#)]
86. Levi, A.; Harris, K.R.; Wechter, W.P.; Kousik, C.S.; Thies, J.A. DNA markers and pollen morphology reveal that *Praecitrullus fistulosus* is more closely related to *Benincasa hispida* than to *Citrullus* spp. *Genet. Resour. Crop Evol.* **2010**, *57*, 1191–1205. [[CrossRef](#)]

87. Rodríguez-Moreno, L.; González, V.M.; Benjak, A.; Martí, M.C.; Puigdomènech, P.; Aranda, M.A.; Garcia-Mas, J. Determination of the melon chloroplast and mitochondrial genome sequences reveals that the largest reported mitochondrial genome in plants contains a significant amount of DNA having a nuclear origin. *BMC Genom.* **2011**, *12*, 424. [[CrossRef](#)] [[PubMed](#)]
88. Heneidak, S.; Khalik, K.A. Seed coat diversity in some tribes of *Cucurbitaceae*: Implications for taxonomy and species identification. *Acta Bot. Bras.* **2015**, *29*, 129–142. [[CrossRef](#)]
89. Gu, C.; Tembrock, L.R.; Zheng, S.; Wu, Z. The complete chloroplast genome of *Catha edulis*: A comparative analysis of genome features with related species. *Int. J. Mol. Sci.* **2018**, *19*, 525. [[CrossRef](#)] [[PubMed](#)]
90. Logacheva, M.D.; Penin, A.; Samigullin, T.H.; Vallejo-Roman, C.M.; Antonov, A.S. Phylogeny of Flowering Plants by the Chloroplast Genome Sequences: In Search of a “Lucky Gene”. *Biochemistry* **2007**, *72*, 1324–1330. [[CrossRef](#)] [[PubMed](#)]
91. Song, W.C.; Ji, C.X.; Chen, Z.M.; Cai, H.H.; Wu, X.M.; Shi, C.; Wang, S. Comparative analysis the complete chloroplast genomes of nine *Musa* Species: Genomic features, comparative analysis, and phylogenetic implications. *Front Plant Sci.* **2022**, *13*, 62. [[CrossRef](#)] [[PubMed](#)]