The complete chloroplast genome sequence of wax gourd (*Benincasa hispida*), and comparative analyses with related species revealing evolutionary dynamics within Benincaseae

Weicai Song
QUST: Qingdao University of Science and Technology
Zimeng Chen
QUST: Qingdao University of Science and Technology
Li He
QUST: Qingdao University of Science and Technology
Feng Qi
QUST: Qingdao University of Science and Technology
Hongrui Zhang
QUST: Qingdao University of Science and Technology
Chao Shi (chsh1111@aliyun.com)
QUST: Qingdao University of Science and Technology
Shuo Wang
QUST: Qingdao University of Science and Technology

Research Article

Keywords: Benincasa hispida, chloroplast genome, repeats, divergence, phylogeny

Posted Date: February 28th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1143844/v1

Abstract

Benincasa hispida (wax gourd) is a crucial important crop of Cucurbitaceae with enormous economic and medicinal importance. Here, we first report the *de novo* assembled and completely annotated the chloroplast genome of wax gourd with 156,758 bp in total that comprises of 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. The chloroplast genome contains 131 genes, including 86 protein-coding genes, 37 tRNAs and 8 rRNAs. Comparison analyses among *B. hispida* and three other species from Benincaseae presented a great extent of conversation regarding nucleotide content, genome structure, codon usage, synonymous and non-synonymous substitutions, putative RNA editing sites, microsatellites, and oligonucleotide repeats. LSC and SSC regions were detected to be much more variety than IR regions through divergent analysis among species within Benincaseae. Notable IR contraction and expansions were observed, causing the difference in genome size and gene duplication, deletion and the presence of pseudogenes. Intronic gene sequences such as *trnR-UCU – atpA, atpH – atpI* were observed as high divergent regions and genes namely *ycf1, accD, ccsA and matK* presented more various than other genes. Two types of phylogenetic analysis based on complete cp genome and 72 genes suggested the sister relationships of *B. hispida* with genus *Citrullus, Lagenaria and Cucumis*. The cp genome of *B. hispida* provide valuable genetic resource for the detection of molecular marker, research of the taxonomic discrepancies and the inference of phylogenetic relationships among species within Cucurbitaceae.

1. Introduction

Cucurbitaceae is a moderately large family of about 130 genera and 900 species (Christenhusz and Byng, 2016). Species of Cucurbitaceae have had a long and intimate association with human beings due to their crucial economic importance in the warmer regions (Yang and Walters, 1992). 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 group of crops in family Cucurbitaceae (Nandecha et al., 2010) As one of them, *Benincasa* represents a monotypic genus with a single species that belongs to tribe Benincaseae (Cucurbitaceae) (Steward, 1969). All of them are economic valuable fruit crops. As the only taxonomic species of *Benincasa*, the wax gourds (*Benincasa hispida*) are not hard to find in the local markets since: 1) it is a rather highly commercialized vegetable due to its long shortage life properties; 2) it bears giant fruit with normally 80 cm in length and weight over 20 kg (Steward, 1969; Xie et al., 2019).

Wax gourds are widely distributed in temperate and sub-temperate climates such as China, Japan, Korea, India and several tropical countries. To date, it is being increasingly popular in Caribbean and the United States. There is no firmly agreement on the origin of wax gourd, while Java and Japan are places that commonly presumed location (Xie et al., 2019). Wax gourd as an important vegetable crop for both nutritional and medical applications (Naik et al., 2016). Its pharmaceutical values cover various aspects, including central nervous system diseases (muscle tension, Alzheimer's disease (Al-Snafi, 2013), gastroprotective diseases (Rachchh and Jain, 2008), depressant-like activities (Dhingra and Joshi, 2012), diabetes, dropsy, diseases related to liver, urinary diseases, and heart diseases. Other effects namely hypolipidemic, antioxidant, anti-inflammatory, antipyretic, anti-angiogenic (Lee et al., 2005) and antimicrobial properties of *B. hispida* are also reported (Daniell et al., 2016; Qadrie et al., 2009). Research has reported that the seed of *B. hispida* contains saponin, urea, citrulline, oleic acid, and fatty acids (Bimakr et al., 2012; Grover et al., 2001).

The chloroplast (cp) is a self-replicating organelle that consists of homogeneous circular DNA molecules. The Double strand DNA inside cp genome ranged from 70 to 520 kb in algae and generally more conserved in land plants that range from 120 to 160 kb. Although the specific nucleotide sequences vary from different species, the quadripartite structure and organization remain a firm consistency, which can be classified into four sections: a large single copy (LSC) region and a small single copy (SSC) region, separated by a pair of inverted repeats (IRa and IRb) (Palmer, 1991). As the metabolic centers, cp genome remains highly conservative to sustain the normal physiological function of cells, especially for genes related to photosynthesis. Despite its conservancy, variations regarding gene content and genome size can be caused by substitutions; insertions and deletions; structural changes such as IR contraction and expansion, genome rearrangements, and translocations

(Ahmed et al., 2012; Sloan et al., 2014). This polymorphism and diversity enable cp genome to perform taxonomic and phylogenetic discrepancies analysis, providing an authentic, time-effective, cost-effective, and also reliable method for population taxonomic and phylogenetic analysis, population genetics studies and evolutionary investigation (Ahmed, 2015; Lössl and Waheed, 2011).

Given wax gourd possesses such giant consumption among Asian communities, yet no full chloroplast genome of *Benincasa* has been provided in any public database to date. The nuclear gene of wax gourd has been reported (Xie et al., 2019) whereas the sequencing of complete nuclear genome restricted to investigate functional genomics because of their large genome size. Here, we first sequenced and assembled the complete chloroplast genome sequence of *B. hispida* and submitted the data on National Center for Biotechnology Information (NCBI). This study is the first comprehensive report of the cp genome of *B. hispida*, and performed comparative analyses with 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) pattens of repeats and microsatellite; 4) high divergent regions; 5) phylogenetic relationships among Cucurbitaceae. The result provided may contribute to unfolding the taxonomical discrepancies, identifying suitable gene markers, and inferring phylogenetics 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, Yunnan Province, 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 (Doyle and Doyle, 1990), and the quantity and quality of extracted DNA was assessed by spectrophotometry and the integrity was evaluated using a 1% (w/v) agarose gel electrophoresis (Wang et al., 2021). The 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 Genomes assembly and annotations

The raw data were preprocessed using Trimmomatic 0.39 software (Bolger et al., 2014), 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 (Andrews, 2010) and MULTIQC software (Ewels et al., 2016), 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 (Guo et al., 2020). Short reads were de novo assembled into long contigs with SOAPdenovo 2.04 (Luo et al., 2012) by setting kmer values of as 35, 44,71 and 101. Finally, the long-contigs complete sequence expansion and gap filling using Geneious ver 8.1 (Muraguri et al., 2020), which forms the complete chloroplast genome. The complete chloroplast genome was further validated and calibrated by using de novo splicing software NOVOplsty 4.2 (Dierckxsens et al., 2017) GeSeq (Tillich et al., 2017) was used to annotate the *de novo* assembled genomes, and tRNAscan-SE ver 1.21 (Lowe and Eddy, 1996) was applied to detect tRNA genes with default settings, and RNAmmer (Lagesen et al., 2007) was used to validate rRNA genes with default settings. As a final check, we compared the results with the reference sequence and correct the misannotated genes by GB2Sequin (Lehwark and Greiner, 2019) in an artificial way. The circular map of the genomes was drawn by using Organellar Genome DRAW (OGDRAW) (Lohse et al., 2007). The newly assembled *B. hispida* chloroplasts genomes were deposited in GenBank with the accession numbers MW362306.

2.3 Chloroplast genomes comparison

In order to gain a better understand of the characters in cp genome of *B. hispida*, we selected three species that not only closely related to *B. hispida* but also representative in Benincaseae to perform comparative analysis. Sequences of their complete chloroplast genome were downloaded from NCBI data-base, with the accession number following: *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 v10.1.3 (Kearse et al.,2012) and relative synonymous codon usage (RSCU) of protein-coding genes was evaluated by MEGA-X (Kumar et al., 2018). We also used Predictive RNA Editors for Plants chloroplast (PREP-cp) (Mower, 2009) were used to investigate putative RNA editing sites in the cp genome of *B. hispida*, *C. colocynthis*, *C. lanatus* and *L. siceraria*.

2.5 Repeat sequences and SSRs analysis

MIcroSAtellite (Misa) (Beier et al., 2017) was used to determine Simple Sequence Repeats (SSRs) or microsatellites in cp genome 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 (Kurtz et al.,2001) 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 (Amiryousefi et al., 2018b) 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 (Frazer et al., 2004)was used to compare the cp genome of four species using Shuffle-LAGAN model with *L. siceraria* set as the reference sequence.

DnaSP was used to perform Slicing Window analysis using multiple alignment of complete cp genome of four selected species. DnaSP 6.12 (Rozas et al., 2017) was also used to determine the synonymous (Ks) and non-synonymous (Ka) substitutions and their ratio (Ka/Ks). Geneious was used to detect the types, numbers, length and position of SNPs and InDels among LSC, SSC and IR regions.

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 (Table S7). Maximum likelihood (ML) tree was constructed through two approaches. One phylogenetic tree was constructed using complete cp genome with both IR regions included and the other was built with 72 gene sequences. MAFFT alignment were made using concatenated 72 gene sequences and the best fit model was found by MEGA-X. All Indels was excluded for both alignments, leaving only substitutions for ML analysis. The best fit models applied for both three were GTR + G, determined based on Bayesian inference (BI) (Zhu et al., 2018).

3. Results

3.1 chloroplast genome assembly, organization, and features of *Benincasa hispida*

The paired-end sequencing of *Benincasa 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 was submitted to NCBI under accession number MW362306 after a throughout check of correctness. As showing in Table 1 and Fig. 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		*Benincasa hispida*
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
	tRNA (%)	53.2
	ALL gene %	39.4
Protein coding part (CDS) (% bp)		51.1
All gene (% bp)		71.6
Non-coding region (% bp)		28.4

The cp genome of *B. hispida* 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 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 content of different gene types, the number of rRNA (55.2%) and tRNA (53.2%) was relatively high, and that of CDSs was 37.9%. In total, 18 genes contained introns, 16 of which (10 protein-coding genes and 6 tRNA genes) contained one intron and two CDSs (*ycf3* and *clpP1*) possessed two introns (Table S1). Among these genes, 17 genes were duplicated in IR regions except one trans-splicing gene, which was observed in *rps12* gene with 5' end located in LSC region and 3' end duplicated in IR regions. The truncation event was observed in *ycf1* gene that started from 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 hispida. The number of asterisks after the gene names indicates the number of introns contained in the genes.

Caregory of Genes	Group of genes	Gene name
photosynthesis related genes	Large subunit of rubisco	*rbcL*
	Photosystem I	*psaA, psaB, psaC, psaI, psaJ*
	Assembly/srability of photosystem I	*ycf3**, ycf4*
	Photosystem II	*psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK,psbL, psbM, psbN, psbT, psbZ*
	ATP synthase	*atpA, atpB, atpE, atpF*, atpH, atpI*
	Cytochrome b6/f complex	*petA, petB*, petD*, petG, petL, petN*
	Cytochrome c synthesis	*ccsA*
	NADH dehydrogenase	*ndhA*, ndhB*, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK*
Transcription and translation related genes	RNA polymerase subunits / transcription	*rpoA, rpoB, rpoC1*, rpoC2*
	Small subunit of ribosomal proteins	*rps11, rps12*(*2), rps14, rps15, rps16*, rps18, rps19, rps2, rps3, rps4, rps7* (*2), *rps8*
	Large subunit of ribosomal proteins	*rpl14, rpl16*, rpl2*(*2), rpl20, rpl22, rpl23*(*2), *rpl32, rpl33, rpl36*
	translation initiation factor	*infA*
RNA genes	ribosomal RNA	*rrn16* (*2), *rrn23* (*2), *rrn4.5* (*2), *rrn5* (*2)
	transfer RNA	*trnA-UGC* (*2), trnR-ACG (*2), trnR-UCU, trnN-GUU (*2), trnD-GUC, trnC-GCA,trnQ-UUG,trnE-UUC, trnG-GCC, trnG-UCC*, trnH-GUG, trnI-CAU (*2), trnI-GAU* (*2), trnL-CAA (*2), trnL-UAA*, trnL-UAG, trnK-UUU*, trnfM-CAU, trnM-CAU, trnF-GAA, trnP-UGG, trnS-GCU, trnS-GGA, trnS-UGA, trnT-GGU, trnT-UGU, trnW-CCA, trnY-GUA, trnV-GAC (*2), trnV-UAC**
Other genes	*RNA processing*	*matK*
	carbon metabolism	*cemA*
	fatty acid synthesis	*accD*
	proteolysis	*clpP1***
	component of TIC complex	*ycf1 (*2)*
	hypothetical proteins	*ycf2 (*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, having 26,703 codons that fit in 64 codon types. The result of amino acid frequency analysis showing that Leucine with 10.5% occurrence was the most abundant amino acid followed by Isoleucine with 8.5%. The number of Cysteine with only 1,1% abundance was the least occurred amino acid.

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

3.3 Putative RNA editing site within Benincaseae

RNA editing event is typical in the cp genome of most land plants and essential in understanding the chloroplast genome at the transcript level. Out of that purpose, we determined the RNA editing site in the cp genome of four species from Benincaseae. In the cp genome of *Benincasa hispida*, PREP-web found 58 putative RNA editing sites in 21 CDS genes (Table S3a). Among these genes, *ndhB* gene with 13 editing sites was determined to be the most variant gene, followed by *ndhD* (8 sites) and *rpoB* (5 sites). We also found that 81% of all RNA editing events occurred at the second nucleotide position of codons while none of these events were located in the third codon position.

Moreover, these RNA editing events result in post-transcriptional substitutions, causing amino acid conversions. In the group of these conversions, fifteen-four out of fifteen six RNA editing sites led to hydrophobic products, comprising Phenylalanine (9), Isoleucine (5), Leucine (32), Methionine (2), Valine (4), Tryptophan (2). Four exceptions led to hydrophilic (neutral) amino acid products, including Cysteine (1), Tyrosine (2), and Serine (1). What is more, Serine to Leucine was found to be the most abundant post-transcriptional substitutions with 41.82% of all RNA editing events, followed by Proline to Leucine (14.55%) and Serine to Phenylalanine (7.27%). Worth mentioning, two RNA editing events were detected that transformed ACG (Thr) to initiation codon AUG, resulting in the start of translation in *ndhB* and *ndhD* gene.

As shown in Table S3b, the total number of RNA editing sites detected was 57 in Citrullus lanatus, 55 in Lagenaria siceraria and Citrullus colocynthis. All 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 cp genome of *Benincasa hispida, Citrullus lanatus, Lagenaria siceraria* and *Citrullus colocynthis* using MISA-web (Beier et al., 2017) and high similarity was revealed among four species. 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 mononucleotide (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 frequency of tetranucleotide and pentanucleotide were only 3% and 0.42%, and that of hexanucleotide repeats was absent in all species (Fig. 2C). Moreover, most of the mononucleotide repeats were A/T motifs while AT/TA motifs comprised 68% of dinucleotide repeats (Table S4).

We also analyzed the distribution of SSRs in two types of different regions, specifying in LSC/IR/SSC regions and intergenic spacer (IGS) /gene regions. According to the result, most of the repeats were located in LSC region, varying from 136 in *C. lanatus* to 148 in *B. hispida*. Second by SSC region (38 in *B. hispida*) and IR regions. Noticeably, the SSC number in IR regions in all species was 26 except *L. siceraria* (24), implementing that IR regions were more conserved than LSC and SSC regions (Fig. 2A). IGS regions were determined to be highly abundant of SSRs in comparison with gene regions. We found 125 SSRs within 46,150 bp IGS regions while 116 SSRs in 112,281 bp gene regions, meaning the density of SSRs in IGS regions was 2.62 times of that of gene regions (Fig. 2B). And similar results were present in all species.

Oligonucleotide repeat sequences were also analyzed using REPuter program (Kurtz et al., 2001). to detect the abundance of four types oligonucleotide repeats, including forward (F), palindromic (P), reverse (R), and complementary (C). Though minor

variations presented about the total number of oligonucleotide repeats, the distribution of four types of repeats and the size of repeats presented an obvious resemblance. In terms of the number of oligonucleotide repeats and its distribution in each type, we found 42 repeats (F = 16, P = 22, R = 4) in 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* (Fig. 2A). The length of repeats was mostly found between 20 to 24 bp (Fig. 2C). The palindromic repeats were the most abundant repeats followed by forward repeats, whereas the number of reverse repeats was few. None of the species had complementary repeats. We also located the region of each oligonucleotide repeats; significant consistency was presented among four species. The number was exactly the same in all species regarding the repeats located in the IRs 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 (5, Fig. 2B).

3.5 IR contraction and expansion

The genome length of chloroplast ranged from 159,758 bp (*B. hispida*) to 157,147 bp (*C. colocynthis*). Besides, in the cp genome of *B. hispida*, the length of IR regions was the shortest with 260,080 bp while that of SSC region was the longest with 180,060 bp (Table S5). Thus, we inferred that the variation in size of cp genome was contributed by the expansion and contraction of IR regions with the evidence followed (Fig. 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* gene and six of which were at J_{SA} while the other two were located in SSC region completely. Moreover, the *ycf1Ψ* (pseudo copy) was only presented in two species (*B. hispida* and *L. siceraria*) at J_{SB} and were 3 bp and 25 bp into 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 few bp into IRb region, except *C. hystrix* with 21 bp and *B. hispida* (completely located in SSC region).

The *rpl2* gene was found close to J_{LB} while that of two species *(C. moschata* and *C. lanatus*) were into LSC region with 11 bp and 6 bp respectively. At the same time, the duplicate *rpl2* gene were absent in the same specific two species. The *rps19* gene was the most variant gene among all genes that close to the IR junction. In four species, *rps19* gene were 2 bp into IRb region and the left 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 percent of identity between sequences and colored regions of high conservation using mVISTA program (Frazer et al., 2004). As showing in Fig. 4, the sequences varied remarkably among different regions. Firstly, most of the differences were located in the LSC and SSC regions while IR regions were almost identical among four species except *rps12* gene, revealing that IR regions were more conserving than LSC and SSC regions. Moreover, IGS regions revealed remarkably divergent than gene regions. Notable divergent non-coding regions including: *trnR-UCU – atpA, atpH – atpI, trnT-GGU – psbD, trnL-UAA –trnF-GAA, accD – pasI, ndhF – rpl32*. While genes such as *ycf1*, *ycf2*, *accD*, *psbA*, *ccsA*, *ndhF* and *matK* were found to be highly divergent coding genes among all.

Ka/Ks ratio is an essential index to identify a mutation from neutral, purifying, and beneficial. Thus, we compared *B. hispida* with *C. colocynthis*, *C. lanatus*, and *L. siceraria* respectively to analyze the synonymous substitutions (Ks), non-synonymous substitutions (Ka) and their ratio (Ka/Ks) of 73 PCGs (Table S6). Among all, 18 genes could not be determined due to absent information (Ks = 0). After deleting these genes as well as non-substitution results, we found that genes carrying out photosynthesis functions revealed Ka/Ks = 0 or at relatively low values, indicating that these groups of genes were fairly conserved. The Ka/Ks ratio of 26 genes was lower than 0.5 and that of 96% genes was lower than one, with only three expectations. The number of *accD* gene were relatively close to one (1.09, 0.85 and 0.92), signifying that the *accD* gene experienced neutral mutation. The Ka/Ks ratio of *rpl36* gene was greater than one with absent information in *L. siceraria*. Outstandingly, the number of *rpl22* gene was the greatest number that up to 2.41 and the ratio between the smallest and Ka/Ks greatest number was 0.5. Thus, we could infer that a beneficial mutation and rapid development had occurred to the *rpl22* gene among four species in Benincaseae.

To get a holistic understanding of sequence divergence, we performed slicing window analysis to visualize the nucleotide variability values of all cp genomes. We found that none of the πvalues of CDS genes exceed 0.05 and the IGSs revealed more

divergent than gene regions, which result was in consistent with the previous analysis mentioned. It can be clearly seen in the figure that SSC and LSC regions were much more divergent than IR regions, the π value of which was remarkably low and mirror symmetrized with SSC as the center (Fig. 5).

3.7 Phylogenetic analysis

To locate the phylogenetic position of *Benincasa hispida* precisely, we selected 26 species and constructed two phylogenetic trees using the complete cp genome (Fig. 6a) and 73 selected CDSs (Fig. 6b) respectively. And the results all supported 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 two approaches presented highly consistency with two main variations. Firstly, in general, the bootstrap values in the tree that applied complete cp genome revealed higher than tthe tree constructed with 73 CDSs (Fig. 6b). In addition, Begoniaceae was a sister group with Coriariaceae and Corynocarpaceae according to Fig. 6a while in Fig. 6b, Coriariaceae and Corynocarpaceae were the early-diverging lineages of Begoniaceae. However, only 82 bootstrap values support the former situation (Fig. 6a) while 94 support the second (Fig. 6b).

4. Discussion

In present study, we sequenced and reported the complete chloroplast genome of *Benincasa hispida* and performed comparative analyses with other three closely related species selected from the *Benincaseae* but relevantly distinct enough to get valuable results, providing the basic genetic data available for phylogeographic and population genetic investigation (Poczai and Hyvönen, 2017; Shaw et al., 2007).

The cp genome revealed high similarities in terms of quadruple structure, gene content, and organization in Benincaseae (Bhowmick and Jha, 2015; Hu et al., 2009) and in other angiosperms (Bausher et al., 2006). The genome size differed less than 400 bp with almost identical gene numbers, signifying that the cp genomes among four analyzed species were firmly conservative on the whole. The GC content of *B. hispida* various among different regions and function. The rRNA sequences were considerably rich in GC bases, as a consequence, IR regions that abundant in rRNA presented higher GC content than other regions. These findings agree with previous studies (Guo et al., 2018; Mehmood et al., 2020).

However, variations still existed which provided valuable information to understand the structure development and evolution (Daniell et al., 2021; Shaw et al., 2007). The bias of codon usage in plant cp genome is an important evolutionary feature for the studies of mRNA translation, new gene discovery, and molecular biology (Yang et al., 2014). Previous studies had confirmed that gene tends to choose preferred synonymous codon for specific amino acid rather than randomly distributed (Li et al., 2019; Sorimachi, 2010), which led us to analyze the codon usage bias to get a thread on laws of genetic evolution. Our study showed that genes in *B. hispida* prefer codons with A/T in the third position, which feature was in consistent with previous studies (Saina et al., 2018; Wang et al., 2017).

Microsatellites or SSRs are widely distributed in cp genomes that serve as molecular markers for phylogenetic relationships inferring (Ahmed et al., 2013; Cavender-Bares et al., 2015). Moreover, SSRs are also related to different types of genome rearrangements, recombination and large inversions (Guisinger et al., 2011; Song et al., 2019). Similar with previous studies, we found that mononucleotide repeats were the most abundant types of repeats and the number of which in LSC region far surpassed that of SSC and IR regions (Jeon and Kim, 2019). What is more, a greater number of palindromic repeats were found among four types of repeats while previous studies revealed that the forward repeats were the most abundant repeats (Abdullah et al., 2019; Saina et al., 2018). We specifically analyzed the abundance of SSRs that differing from gene regions to intronic gene regions and verified that IGSs contained much higher SSRs density than the other. Thus, we inferred that IGS regions might experience more gene rearrangements and recombination than gene regions. Moreover, our result support the hypothesis that cpSSRs are more often composed by polyA or polyT rather than polyG or polyC (Raubeson et al., 2007; Shen et al., 2017), implicating that IGSs might be relatively rapidly mutating regions (Liu et al., 2018; Provan et al., 2001).

It is commonly agreed that the variation of genome size in the chloroplast is the consequence of IR contraction and expansion, leading to gene duplication, deletion, and the presence of pseudogenes (Abdullah et al., 2019; Zhu et al., 2021). We found that the *ycf1Ψ* pseudogenes were only detected in *B. hispida* and *L. siceraria*, which were also sister groups in ML phylogenetic trees. What is more, none *rps19Ψ* pseudogene was observed in all species analyzed, whose presence was thought to be the cause of loss of functional ability of *rps19* gene (Menezes et al., 2018; Shahzadi et al., 2020). This result implies that the gene variation at IR boundaries may contribute to understand the cp genome at molecular level as an index for evolutionary investigation (Jansen et al., 2011; Nazareno et al., 2015).

The gene diversity among four Benincaseae species is worth to investigate since the chloroplast genome plays vital role in the study of phylogenetic development, gene flow between species, and population genetics among different species (Cavender-Bares et al., 2015; Li et al., 2018). The non-coding regions were generally agreed to be more conserved than coding regions. And some of the coding genes namely *ycf1, ycf2, matK, accD* and *ndhF* genes were commonly found relatively divergent than others (Amiryousefi et al., 2018a; Du et al., 2017). In addition, LSC and SSC regions was further confirmed to be more divergent in comparison with IR regions (Huo et al., 2019). We also discovered that genes related to photosynthesis with low Ka/Ks ratio showed slow evolution rates and genes such as *ycf1* revealed high evolutional rates, indicating that genes carrying out vital functions were conserved and vice versa (Menezes et al., 2018; Zheng et al., 2017).

To date, protein-coding genes were commonly implemented for phylogenetic tree building (Cui et al., 2019). While the complete cp genome that contains richer information but requires longer time to perform, higher-end equipment and the population distance may be exaggerated for the highly divergence feature of IGS genes (Cheng et al., 2020). In this study, we carried out both methods to build the phylogenetic tree. The tree built with complete cp genome revealed higher bootstrap values in general while the other tree built with coding genes presented slightly phylogenetic order of four species out of twenty-six in total. In general, the phylogenetic position revealed was in consistent with former studies (Heneidak and Khalik, 2015; Levi et al., 2010; Rodríguez-Moreno et al., 2011).

5. Conclusion

In conclusion, our study firstly shed light on the structure and content of cp genome of *B. hispida*, an important economic fruit crop within Asia and several tropical countries. The chloroplast genome of wax gourd with 156,758 bp in total that comprises of 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. The chloroplast genome contains 131 genes, including 86 protein-coding genes, 37 tRNAs and 8 rRNAs. We also offered information regarding the similarities and divergence, include conversation regarding nucleotide content, genome structure, codon usage, synonymous and non-synonymous substitutions, putative RNA editing sites, microsatellites, and oligonucleotide repeats, enriching the understanding of species in Benincaseae. Moreover, the information about highly polymorphic regions was provided as well regarding molecular markers and highly divergent regions that might be useful for further studies of taxonomy and phylogeographic in tribe Benincaseae.

Declarations

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).

CRediT authorship contribution statement

Weicai Song: Software, Formal analysis, Investigation, Visualization, Writing original draft. **Zimeng Chen**: Software, Formal analysis, Visualization, Writing. **Li He**: Investigation, Visualization. **Qi Feng** & **Hongrui Zhang**: Writing - review & editing. **Chao Shi**: Investigation, Writing - review & editing, Funding acquisition. **Shuo Wang**: Conceptualization, Writing - review & editing, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Abdullah., Shahzadi I, Mehmood F, Ali Z, Malik MS, Waseem S, Mirza B, Ahmed I, Waheed MT (2019) Comparative analyses of chloroplast genomes among three *Firmiana* species: Identification of mutational hotspots and phylogenetic relationship with other species of Malvaceae. Plant Gene 19:100199. https://doi.org/10.1016/j.plgene.2019.100199
2. Ahmed I (2015) Chloroplast genome sequencing: Some reflections. J. Next Gener. Seq Appl 2:2469–9853. https://doi.org/10.4172/2469-9853.1000119
3. Ahmed I, Biggs PJ, Matthews PJ, Collins LJ, Hendy MD, Lockhart PJ (2012) Mutational dynamics of *Aroid* chloroplast genomes. Genome Biol Evol 4:1316–1323. https://doi.org/10.1093/gbe/evs110
4. Ahmed I, Matthews PJ, Biggs PJ, Naeem M, Mclenachan PA, Lockhart PJ (2013) Identification of chloroplast genome loci suitable for high-resolution phylogeographic studies of *Colocasia esculenta* (L.) Schott (Araceae) and closely related taxa. Mol Ecol Resour 13:929–937. https://doi.org/10.1111/1755-0998.12128
5. Al-Snafi AE (2013) The Pharmacological Importance of *Benincasa hispida*. A review. Int J Pharma Sci Res 4:0975–9492. https://www.researchgate.net/publication/313676687_The_Pharmacological_Importance_of_Benincasa_hispida_A_review
6. Amiryousefi A, Hyvönen J, Poczai P (2018a) The chloroplast genome sequence of bittersweet (*Solanum dulcamara*): Plastid genome structure evolution in Solanaceae. PLoS ONE 13:1–23. https://doi.org/10.1371/journal.pone.0196069
7. Amiryousefi A, Hyvönen J, Poczai P (2018b) IRscope: an online program to visualize the junction sites of chloroplast genomes. Bioinformatics 34:3030–3031. https://doi.org/10.1093/bioinformatics/bty220
8. Andrews S (2010) FastQC: a quality control tool for high throughput sequence data.Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
9. Bausher MG, Singh ND, Lee SB, Jansen RK, Daniell H (2006) The complete chloroplast genome sequence of *Citrus sinensis* (L.) Osbeck var 'Ridge Pineapple': organization and phylogenetic relationships to other angiosperms. BMC Plant Biol 6:21. https://doi.org/10.1186/1471-2229-6-21
10. Beier S, Thiel T, Münch T, Scholz U, Mascher M (2017) MISA-web: A web server for microsatellite prediction. Bioinformatics 33:2583–2585. https://doi.org/10.1093/bioinformatics/btx198
11. Bhowmick BK, Jha S (2015) Differential heterochromatin distribution, flow cytometric genome size and meiotic behavior of chromosomes in three Cucurbitaceae species. Sci Hortic (Amsterdam) 193:322–329. https://doi.org/10.1016/j.scienta.2015.07.006
12. Bimakr M, Rahman RA, Taip FS, Adzahan NM, Sarker I, Ganjloo MZ, A (2012) 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 17:11748–11762. https://doi.org/10.3390/molecules171011748
13. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170
14. Cavender-Bares J, González-Rodríguez A, Eaton DAR, Hipp AAL, Beulke A, Manos PS (2015) Phylogeny and biogeography of the american live oaks (Quercus subsection Virentes): A genomic and population genetics approach. Mol Ecol 24:3668–3687. https://doi.org/10.1111/mec.13269
15. Cheng Y, Zhang LM, Qi J, Zhang LW (2020) Complete chloroplast genome sequence of *Hibiscus cannabinus* and comparative analysis of the Malvaceae Family. Front Genet 11:277. https://doi.org/10.3389/fgene.2020.00227
16. Christenhusz MJM, Byng JW (2016) The number of known plants species in the world and its annual increase. Phytotaxa 261:201–217. https://doi.org/10.11646/phytotaxa.261.3.1

17. Cui Y, Nie L, Sun W, Xu Z, Wang Y, Yu J, Song J, Yao H (2019) Comparative and phylogenetic analyses of ginger (*Zingiber officinale*) in the family Zingiberaceae based on the complete chloroplast genome. Plants 8:283. https://doi.org/10.3390/plants8080283
18. Daniell H, Jin S, Zhu XG, Gitzendanner MA, Soltis DE, Soltis PS (2021) Green giant—a tiny chloroplast genome with mighty power to produce high-value proteins: history and phylogeny. Plant Biotechnol J 19:430–447. https://doi.org/10.1111/pbi.13556
19. Daniell H, Lin CS, Yu M, Chang WJ (2016) Chloroplast genomes: Diversity, evolution, and applications in genetic engineering. Genome Biol 17:134. https://doi.org/10.1186/s13059-016-1004-2
20. Dhingra D, Joshi P (2012) Antidepressant-like activity of *Benincasa hispida* fruits in mice: Possible involvement of monoaminergic and GABAergic systems. J Pharmacol Pharmacother 3:60–62. https://doi.org/10.4103/0976-500X.92521
21. Dierckxsens N, Mardulyn P, Smits G, NOVOPlasty (2017) : De novo assembly of organelle genomes from whole genome data.Nucleic Acids Res.45,18. https://doi.org/10.1093/nar/gkw955.
22. Doyle JJ, Doyle JL (1990) Isolation of plant DNA from fresh tissue. Focus 12:13–15. https://worldveg.tind.io/record/33886
23. Du YP, Bi Y, Yang FP, Zhang MF, Chen XQ, Xue J, Zhang XH (2017) Complete chloroplast genome sequences of *Lilium*: Insights into evolutionary dynamics and phylogenetic analyses. Sci Rep 7:1–10. https://doi.org/10.1038/s41598-017-06210-2
24. Ewels P, Magnusson M, Lundin S, Käller M (2016) MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32:3047–3048. https://doi.org/10.1093/bioinformatics/btw354
25. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I (2004) VISTA: Computational tools for comparative genomics. Nucleic Acids Res 32:W273–W279. https://doi.org/10.1093/nar/gkh458
26. Grover JK, Adiga G, Vats V, Rathi SS (2001) Extracts of *Benincasa hispida* prevent development of experimental ulcers. J Ethnopharmacol 78:159–164. https://doi.org/10.1016/S0378-8741(01)00334-8
27. Guisinger MM, Kuehl JV, Boore JL, Jansen RK (2011) Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: Rearrangements, repeats, and codon usage. Mol Biol Evol 28:583–600. https://doi.org/10.1093/molbev/msq229
28. Guo L, Guo S, Xu J, He L, Carlson JE, Hou X (2020) Phylogenetic analysis based on chloroplast genome uncover evolutionary relationship of all the nine species and six cultivars of tree peony. Ind Crops Prod 153:112567. https://doi.org/10.1016/j.indcrop.2020.112567
29. Guo S, Guo L, Zhao W, Xu J, Li Y, Zhang X, Shen X, Wu M, Hou X (2018) Complete chloroplast genome sequence and phylogenetic analysis of *Paeonia ostii*. Molecules 23:1–14. https://doi.org/10.3390/molecules23020246
30. Heneidak S, Khalik KA (2015) Seed coat diversity in some tribes of Cucurbitaceae: Implications for taxonomy and species identification. Acta Bot Brasilica 29:129–142. https://doi.org/10.1590/0102-33062014abb3705
31. Hu JB, Zhou XY, Li JW (2009) Development of novel chloroplast microsatellite markers for *Cucumis* from sequence database. Biol Plant 53:793–796. https://doi.org/10.1007/s10535-009-0146-4
32. Huo YM, Gao LM, Liu BJ, Yang YY, Kong SP, Sun YQ, Yang YH, Wu X (2019) Complete chloroplast genome sequences of four *Allium* species: comparative and phylogenetic analyses. Sci Rep 9:1–14. https://doi.org/10.1038/s41598-019-48708-x
33. Jansen RK, Saski C, Lee SB, Hansen AK, Daniell H (2011) 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 28:835–847. https://doi.org/10.1093/molbev/msq261
34. Jeon JH, Kim SC (2019) Comparative analysis of the complete chloroplast genome sequences of three closely related east-asian wild roses (*Rosa* sect. synstylae; Rosaceae). Genes (Basel). 10, 6–8. https://doi.org/10.3390/genes10010023
35. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A (2012) Geneious Basic: Bioinformatics software for sequence data analysis.

https://www.geneious.com/

36. Kumar S, Stecher G, Li M, Knyaz C, Tamura K (2018) MEGA-X: Molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 35:1547–1549. https://doi.org/10.1093/molbev/msy096
37. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. Nucl Acids Res 29:4633–4642. https://doi.org/10.1093/nar/29.22.4633
38. Lagesen K, Hallin P, Rødland EA, Stærfeldt HH, Rognes T, Ussery DW (2007) RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 35:3100–3108. https://doi.org/10.1093/nar/gkm160
39. Lee KH, Choi HR, Kim CH (2005) Anti-angiogenic effect of the seed extract of *Benincasa hispida* Cogniaux. J Ethnopharmacol 97:509–513. https://doi.org/10.1016/j.jep.2004.12.008
40. Lehwark P, Greiner S (2019) GB2sequin - A file converter preparing custom GenBank files for database submission. Genomics 111:759–761. https://doi.org/10.1016/j.ygeno.2018.05.003
41. Levi A, Harris KR, Wechter WP, Kousik CS, Thies JA (2010) DNA markers and pollen morphology reveal that Praecitrullus fistulosus is more closely related to *Benincasa hispida* than to *Citrullus* spp. Genet Resour Crop Evol 57:1191–1205. https://doi.org/10.1007/s10722-010-9559-3
42. Li W, Zhang C, Guo X, Liu Q, Wang K (2019) Complete chloroplast genome of *Camellia japonica* genome structures, comparative and phylogenetic analysis. PLoS ONE 14:1–18. https://doi.org/10.1371/journal.pone.0216645
43. Li X, Li Y, Zang M, Li M, Fang Y (2018) Complete chloroplast genome sequence and phylogenetic analysis of *Quercus acutissima*. Int J Mol Sci 19:1–17. https://doi.org/10.3390/ijms19082443
44. Liu L, Wang Y, He P, Li P, Lee J, Soltis DE, Fu C (2018) Chloroplast genome analyses and genomic resource development for epilithic sister genera *Oresitrophe* and *Mukdenia* (Saxifragaceae), using genome skimming data. BMC Genomics 19:1–17. https://doi.org/10.1186/s12864-018-4633-x
45. Lohse M, Drechsel O, Bock R (2007) OrganellarGenomeDRAW (OGDRAW): A tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. Curr Genet 52:267–274. https://doi.org/10.1007/s00294-007-0161-y
46. Lössl AG, Waheed MT (2011) Chloroplast-derived vaccines against human diseases: Achievements, challenges and scopes. Plant Biotechnol J 9:527–539. https://doi.org/10.1111/j.1467-7652.2011.00615.x
47. Lowe TM, Eddy SR (1996) TRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964. https://doi.org/10.1093/nar/25.5.0955
48. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu YJ, Tang J, Wu G, Zhang H, Shi Y, Liu, Yong, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu SM, Peng S, Liu XQZ, Liao G, Li X, Yang Y, Wang H, Lam J, Wang TW, J (2012) SOAPdenovo2: An empirically improved memory-efficient short-read *de novo* assembler. Gigascience 1, 2047–217X-1-18. https://doi.org/10.1186/2047-217X-1-18
49. Mehmood F, Abdullah, Shahzadi I, Ahmed I, Waheed MT, Mirza B (2020) Characterization of *Withania somnifera* chloroplast genome and its comparison with other selected species of Solanaceae. Genomics 112:1522–1530. https://doi.org/10.1016/j.ygeno.2019.08.024
50. Menezes APA, Resende-Moreira LC, Buzatti RSO, Nazareno AG, Carlsen M, Lobo FP, Kalapothakis E, Lovato MB (2018) Chloroplast genomes of *Byrsonima* species (Malpighiaceae): Comparative analysis and screening of high divergence sequences. Sci Rep 8:2210. https://doi.org/10.1038/s41598-018-20189-4
51. Mower JP (2009) The PREP suite: Predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments. Nucleic Acids Res 37:W253–W259. https://doi.org/10.1093/nar/gkp337
52. Muraguri S, Xu W, Chapman M, Muchugi A, Oluwaniyi A, Oyebanji O, Liu A (2020) Intraspecific variation within Castor bean (*Ricinus communis* L.) based on chloroplast genomes. Ind Crops Prod 155:112779. https://doi.org/10.1016/j.indcrop.2020.112779
53. Naik R, Buha M, Acharya R, Borkar SD (2016) Role of vegetables (*Shaka Dravyas*) in prevention and management of gastro - intestinal tract diseases: A critical review. J Res Tradit Med 2:103–112. https://doi.org/10.21276/jrtm.2016/174

54. Nandecha C, Nahata A, Dixit VK (2010) Effect of *Benincasa hispida* fruits on testosterone-induced prostatic hypertrophy in albino rats. Curr Ther Res 71:331–343. https://doi.org/10.1016/j.curtheres.2010.10.006
55. Nazareno AG, Carlsen M, Lohmann LG (2015) Complete chloroplast genome of *Tanaecium tetragonolobum*: The first Bignoniaceae plastome. PLoS ONE 10:e0129930. https://doi.org/10.1371/journal.pone.0129930
56. Palmer JD (1991) Plastid chromosomes: structure and evolution. Mol Biology Plastids 5–53. https://doi.org/10.1016/b978-0-12-715007-9.50009-8
57. Poczai P, Hyvönen J (2017) The complete chloroplast genome sequence of the CAM epiphyte Spanish moss (*Tillandsia usneoides*, Bromeliaceae) and its comparative analysis. PLoS ONE 12:e0187199. https://doi.org/10.1371/journal.pone.0187199
58. Provan J, Powell W, Hollingsworth PM (2001) Chloroplast microsatellites: New tools for studies in plant ecology and evolution. Trends Ecol Evol 16:142–147. https://doi.org/10.1016/S0169-5347(00)02097-8
59. Qadrie ZL, Hawisa NT, Khan MWA, Samuel M, Anandan R (2009) Antinociceptive and anti-pyretic activity of *Benincasa hispida* (Thunb.) Cogn. In wistar albino rats. Pak J Pharm Sci 22:287–290. https://doi.org/10.0000/PMID19553176
60. Rachchh M, Jain S (2008) Gastroprotective effect of *Benincasa hispida* fruit extract. Indian J Pharmacol 40:271–275. https://doi.org/10.4103/0253-7613.45154
61. Raubeson LA, Peery R, Chumley TW, Dziubek C, Fourcade HM, Boore JL, Jansen RK (2007) Comparative chloroplast genomics: Analyses including new sequences from the angiosperms *Nuphar advena* and *Ranunculus macranthus*. BMC Genomics 8:174. https://doi.org/10.1186/1471-2164-8-174
62. Rodríguez-Moreno L, González VM, Benjak A, Martí MC, Puigdomènech P, Aranda MA, Garcia-Mas J (2011) 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 Genomics 12:424. https://doi.org/10.1186/1471-2164-12-424
63. Rozas J, Ferrer-Mata A, Sanchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, Sanchez-Gracia A (2017) Mol Biol Evol 34:3299–3302. https://doi.org/10.1093/molbev/msx248. DnaSP 6: DNA sequence polymorphism analysis of large data sets
64. Saina JK, Li ZZ, Gichira AW, Liao YY (2018) The complete chloroplast genome sequence of tree of heaven (*Ailanthus altissima* (mill.) (sapindales: Simaroubaceae), an important pantropical tree. Int. J. Mol. Sci. 19, 292. https://doi.org/10.3390/ijms19040929
65. Shahzadi I, Abdullah, Mehmood F, Ali Z, Ahmed I, Mirza B (2020) Chloroplast genome sequences of *Artemisia maritima* and *Artemisia absinthium*: Comparative analyses, mutational hotspots in genus *Artemisia* and phylogeny in family Asteraceae. Genomics 112:1454–1463. https://doi.org/10.1016/j.ygeno.2019.08.016
66. Shaw J, Lickey EB, Schilling EE, Small RL (2007) Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: The Tortoise and the hare III. Am J Bot 94:275–288. https://doi.org/10.3732/ajb.94.3.275
67. Shen X, Wu M, Liao B, Liu Z, Bai R, Xiao S, Li X, Zhang B, Xu J, Chen S (2017) Complete chloroplast genome sequence and phylogenetic analysis of the medicinal plant *Artemisia annua*. Molecules 22:1330. https://doi.org/10.3390/molecules22081330
68. Sloan DB, Triant DA, Forrester NJ, Bergner LM, Wu M, Taylor DR (2014) A recurring syndrome of accelerated plastid genome evolution in the angiosperm tribe Sileneae (Caryophyllaceae). Mol Phylogenet Evol 72:82–89. https://doi.org/10.1016/j.ympev.2013.12.004
69. Song Y, Zhang Y, Xu J, Li W, Li MF (2019) Characterization of the complete chloroplast genome sequence of *Dalbergia* species and its phylogenetic implications. Sci Rep 9:1–10. https://doi.org/10.1038/s41598-019-56727-x
70. Sorimachi K (2010) Codon evolution in double-stranded organelle DNA: strong regulation of homonucleotides and their analog alternations. Nat Sci 2:846–854. https://doi.org/10.4236/ns.2010.28106

71. Steward FC (1969) Some economic plants: tropical crops: dicotyledons. J. W. Purseglove. Wiley, New York, 1968. 2 vols., xx + 719 pp., illus. $8.50 each. Science 163, 1050–1051. https://doi.org/10.1126/science.163.3871.1050
72. Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, Greiner S (2017) GeSeq - Versatile and accurate annotation of organelle genomes. Nucleic Acids Res 45:W6–W11. https://doi.org/10.1093/nar/gkx391
73. Wang W, Yu H, Wang JH, Lei W, Gao J, Qiu X, Wang JS (2017) The complete chloroplast genome sequences of the medicinal plant *Forsythia suspensa* (Oleaceae). Int J Mol Sci 18:2288. https://doi.org/10.3390/ijms18112288
74. Wang Y, Wang S, Liu Y, Yuan Q, Sun J, Guo L (2021) Chloroplast genome variation and phylogenetic relationships of *Atractylodes* species. BMC Genomics 22:1–12. https://doi.org/10.1186/s12864-021-07394-8
75. Xie D, Xu Y, Wang J, Liu W, Zhou Q, Luo S, Huang W, He X, Li Q, Peng Q, Yang X, Yuan J, Yu J, Wang X, Lucas WJ, Huang S, Jiang B, Zhang Z (2019) The wax gourd genomes offer insights into the genetic diversity and ancestral cucurbit karyotype. Nat Commun 10:5158. https://doi.org/10.1038/s41467-019-13185-3
76. Yang SL, Walters TW (1992) Ethnobotany and the economic role of the Cucurbitaceae of China. Econ Bot 46:349–367. https://doi.org/10.1007/BF02866506
77. Yang X, Luo X, Cai X (2014) Analysis of codon usage pattern in *Taenia saginata* based on a transcriptome dataset. Parasites and Vectors 7:527. https://doi.org/10.1186/s13071-014-0527-1
78. Zheng XM, Wang JR, Feng L, Liu S, Pang HB, Qi L, Sun LJ, Qiao Y, Zhang WH, Cheng LF, Yang YL, Q.W (2017) Inferring the evolutionary mechanism of the chloroplast genome size by comparing whole-chloroplast genome sequences in seed plants. Sci Rep 7:1–10. https://doi.org/10.1038/s41598-017-01518-5
79. Zhu B, Qian F, Hou Y, Yang W, Cai M, Wu X (2021) Complete chloroplast genome features and phylogenetic analysis of *Eruca sativa* (Brassicaceae). PLoS ONE 16:1–19. https://doi.org/10.1371/journal.pone.0248556
80. Zhu J, Wen D, Yu Y, Meudt HM, Nakhleh L (2018) Bayesian inference of phylogenetic networks from bi-allelic genetic markers. PLoS Comput Biol 14:e1005932. https://doi.org/10.1371/journal.pcbi.1005932

Figures

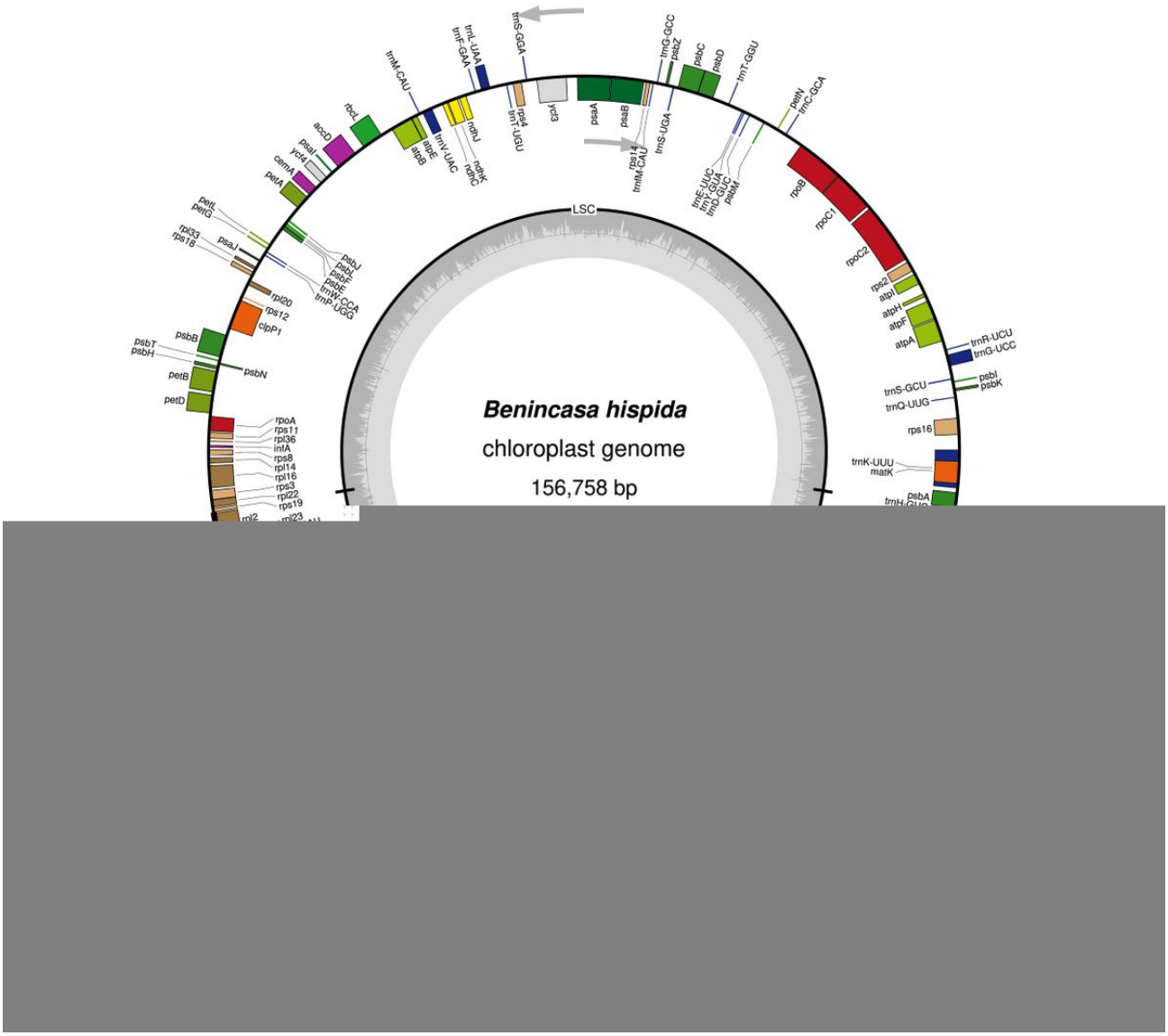

Figure 1

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 were colored based on their function. The borders of chloroplast genome are defined with LSC, SSR, IRa, and IRb. The dashed grey color of inner circle shows the GC content whereas the lighter grey color shows AT content. Asterisks mark genes that have introns.

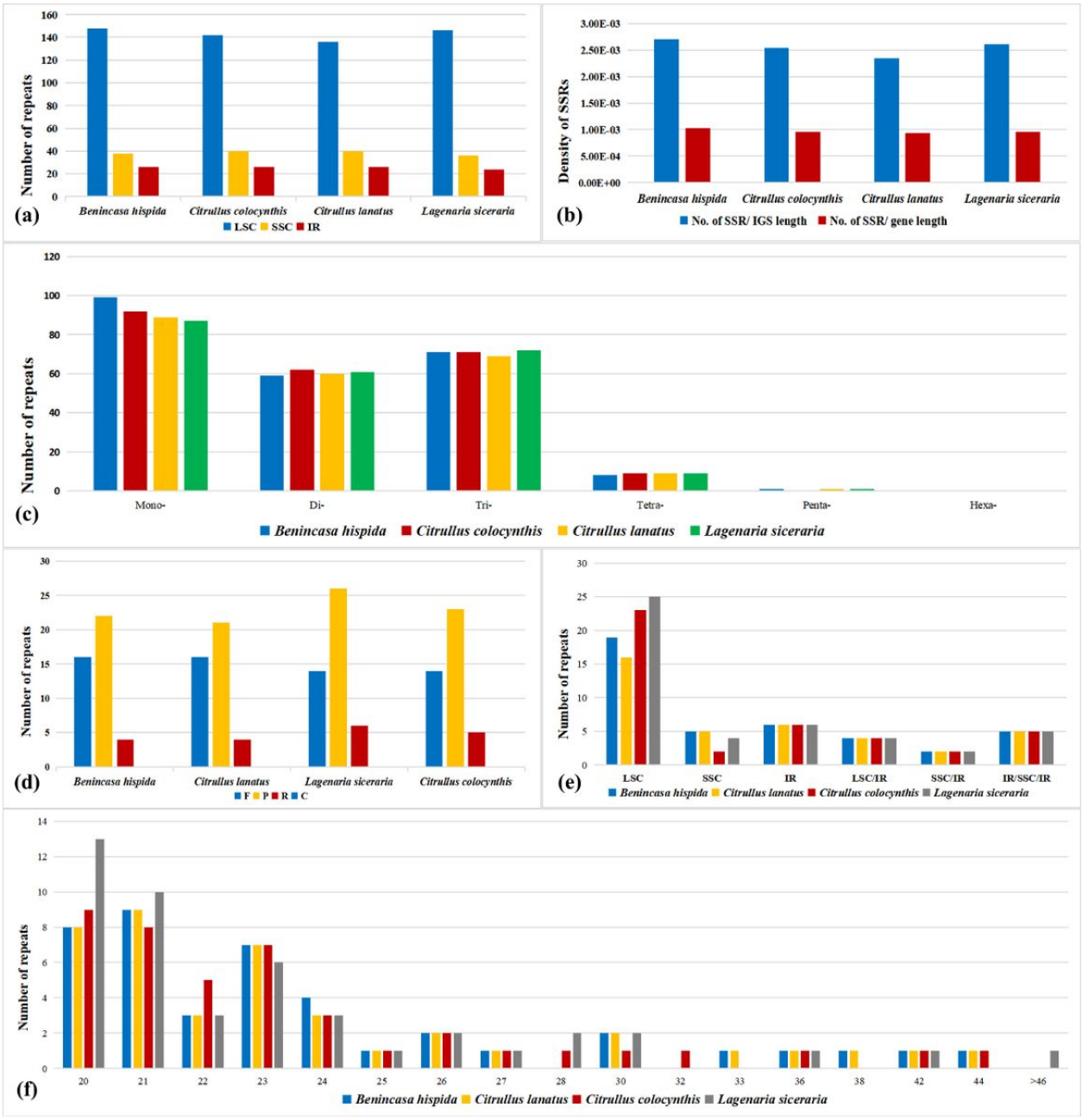

Figure 2

Comparison of microsatellites and oligonucleotide repeats in the chloroplast genomes of Benincaseae species. (a) The number of SSRs in the three main region of chloroplast genome. LSC: large single copy region, SSC: small single copy region, IR: inverted repeat region. (b) The density of SSRs in IGSs (intra gene sequences) and gene regions. (c) The number of different types of SSRs. Mono- represent mononucleotide SSRs, Di- represent dinucleotide SSRs and so on. (d) Different types of oligonucleotide repeats. 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: repeats sequences crossed LSC and IR regions and so on. (f) The number of repeats in different repeats units.

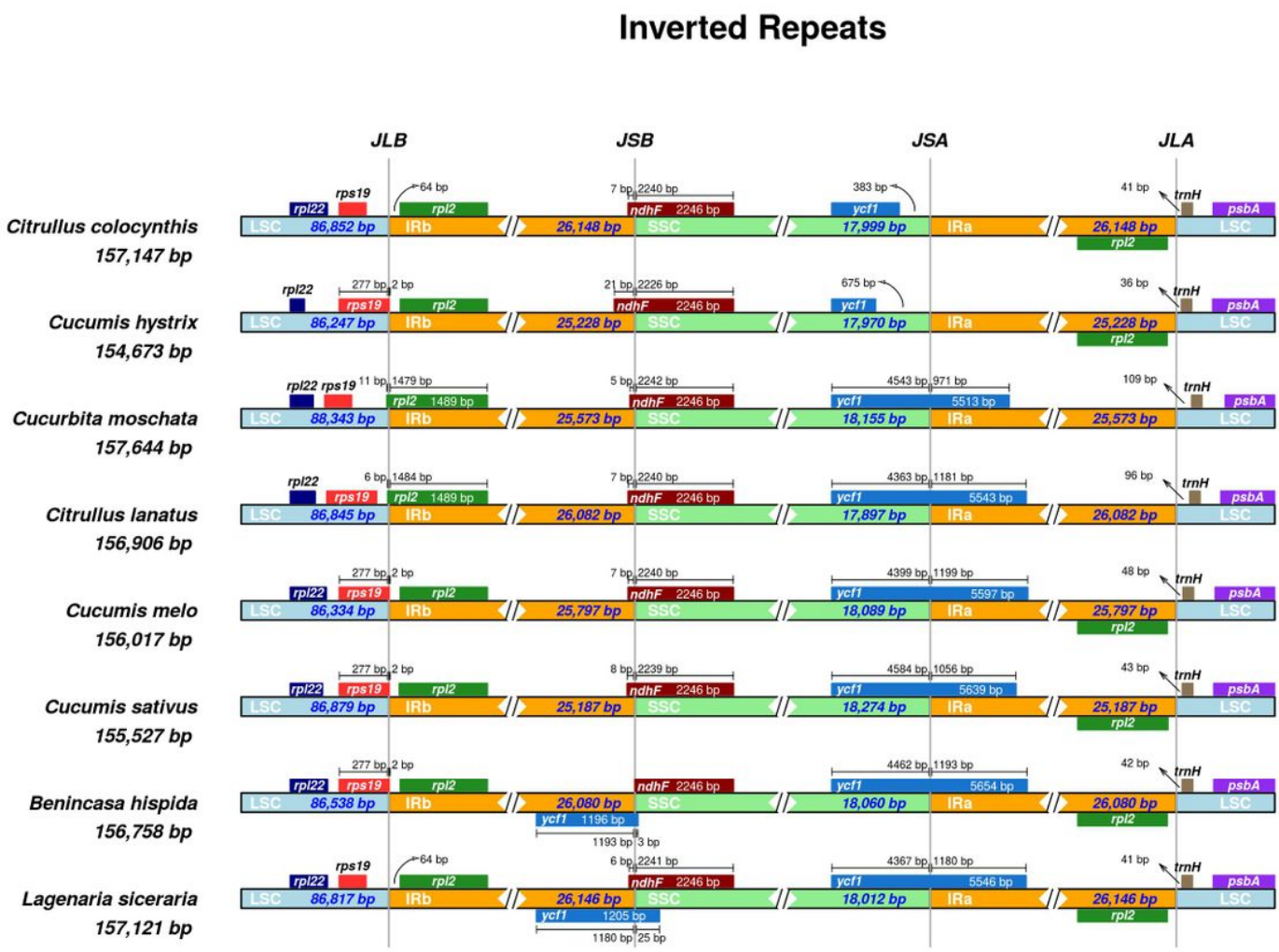

Figure 3

Comparison of junctions between the LSC, SSC, and IRs among eight species. Number above indicates the distance in bp between the ends of genes and the borders sites (distances are not to scale in this figure). The ψ symbol represents pseudogenes. J_{LB} (IRb /LSC), J_{SA} (SSC/IRa), J_{SB} (IRb/SSC), and J_{LA} (IRa/LSC) indicate the junction sites between the quadripartite regions of the genome.

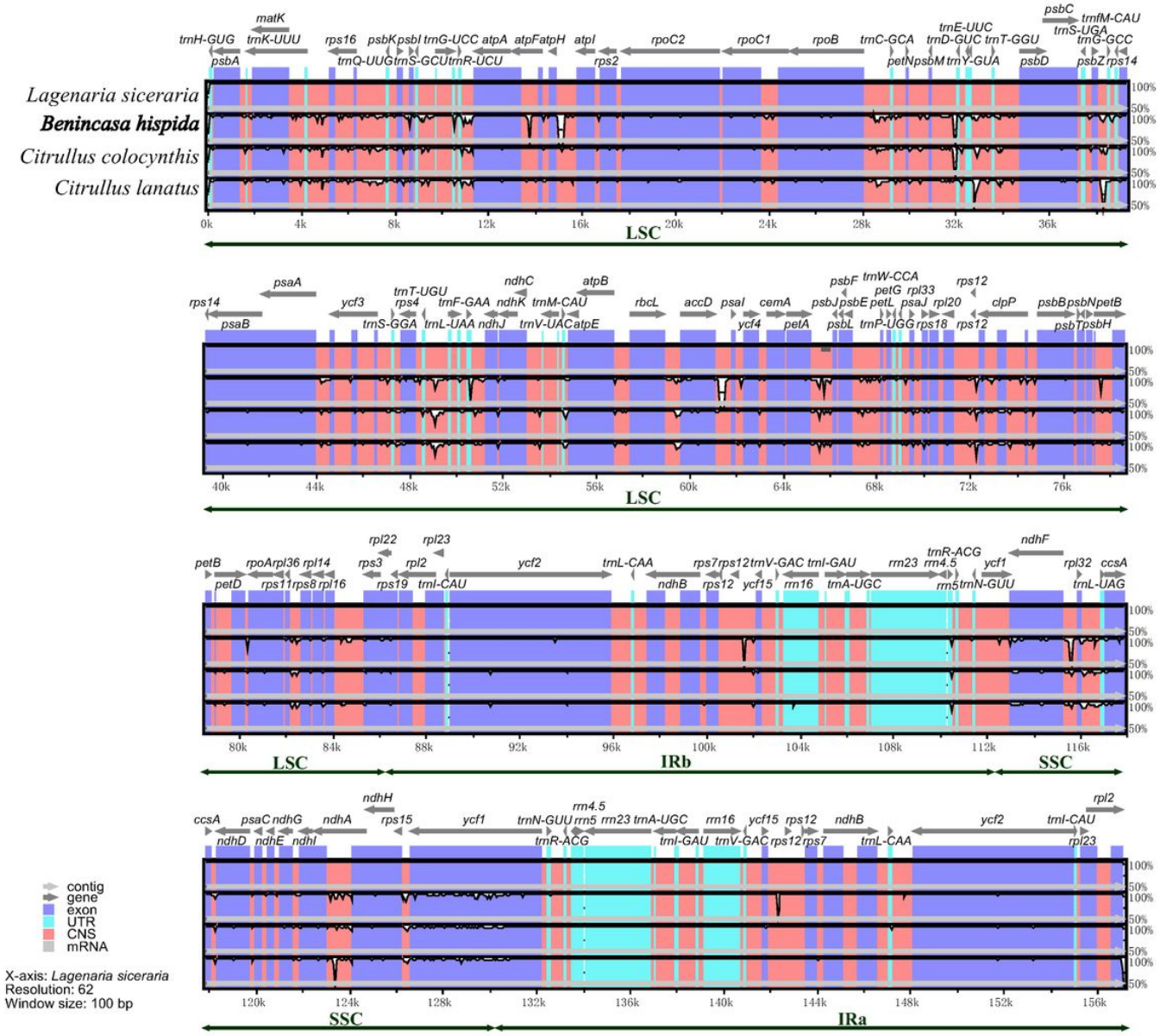

Figure 4

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%).

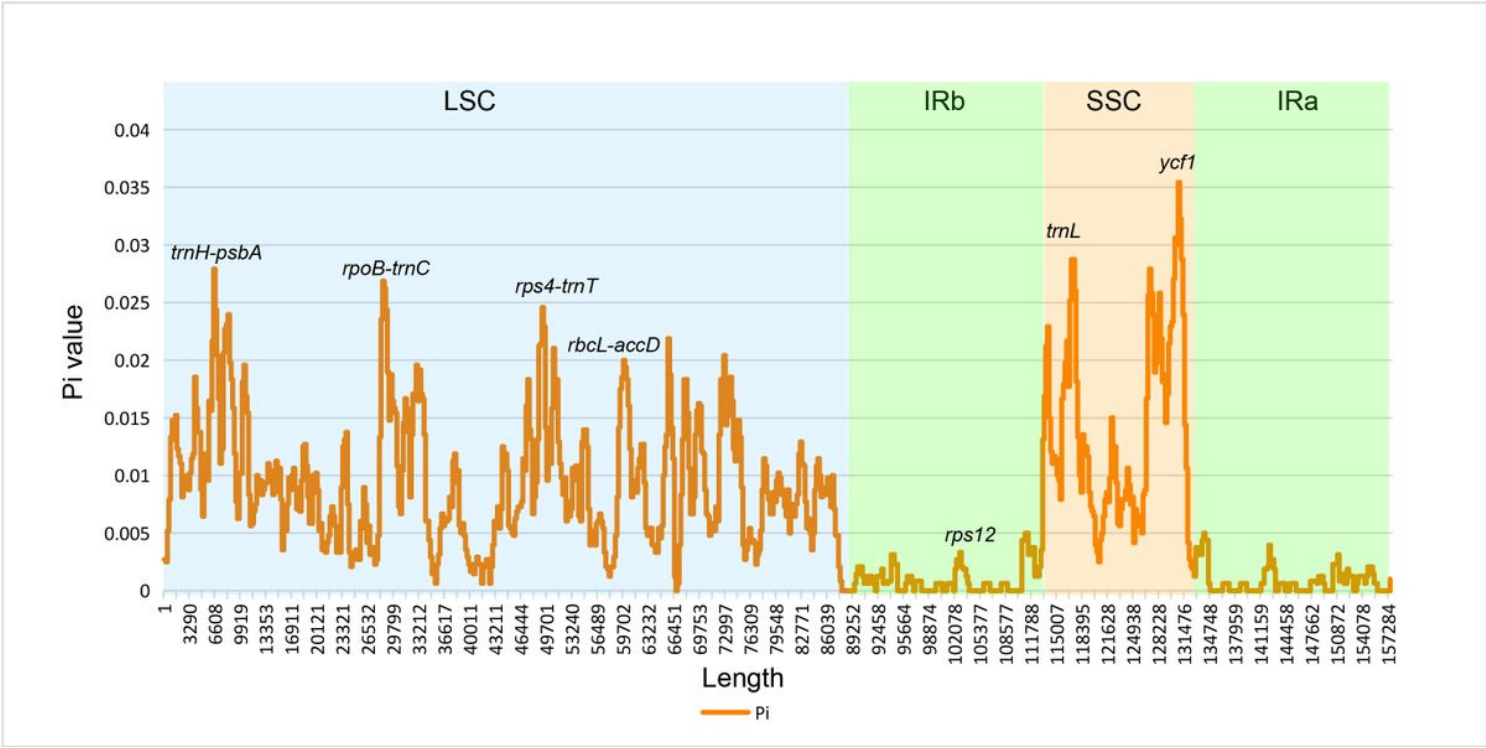

Figure 5

Nucleotide diversity (π) values among the Benincaseae species.

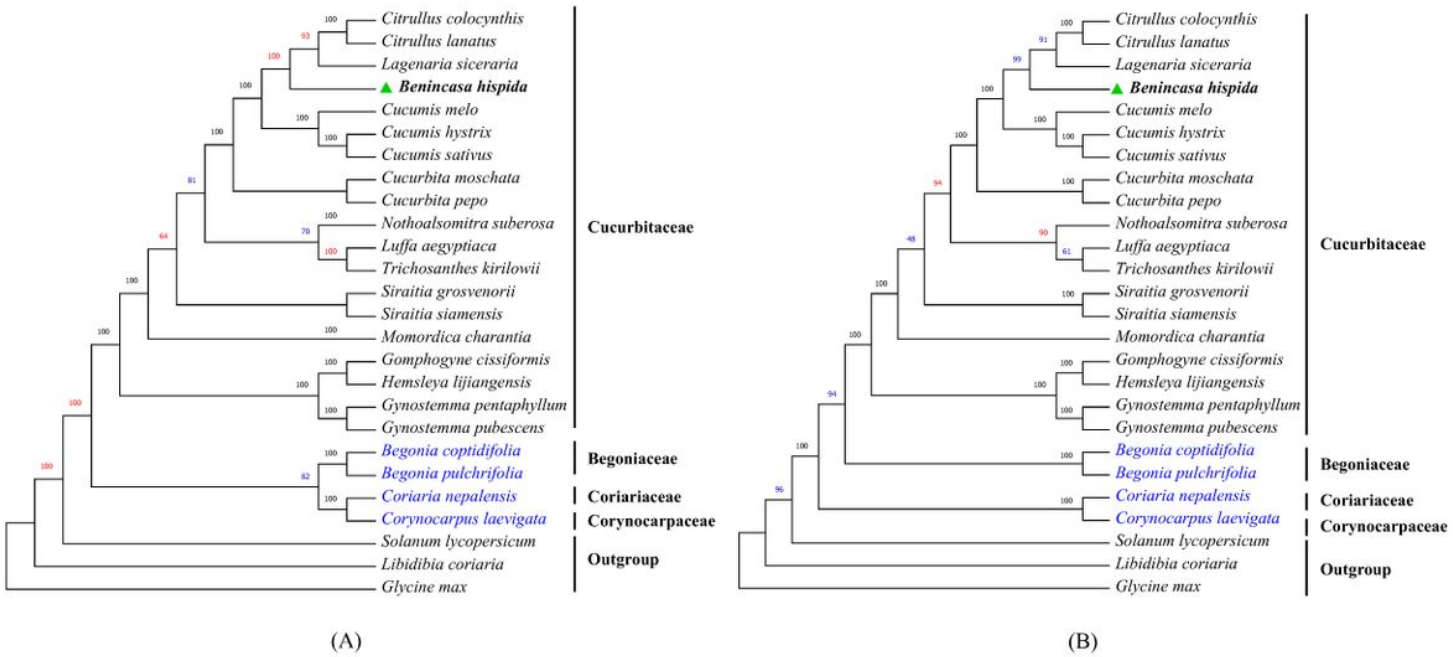

Figure 6

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

Supplementary Files

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

- TableS1intron.xlsx
- TableS2Codon.xlsx
- TableS3RNAediting.xlsx
- TableS4SSRdetails.xlsx
- TableS5regionlength.xlsx
- TableS6KaKs.xlsx
- TableS7tree.xlsx