molecules

MDPI

Article

Comparative Analyses of Chloroplast Genomes of Cucurbitaceae Species: Lights into Selective Pressures and Phylogenetic Relationships

Xiao Zhang [1], Tao Zhou [2], Jia Yang [1], Jingjing Sun [1], Miaomiao Ju [1], Yuemei Zhao [3] and Guifang Zhao [1,*]

[1] Key Laboratory of Resource Biology and Biotechnology in Western China (Ministry of Education), College of Life Sciences, Northwest University, Xi'an 710069, China; zhxiaao@163.com (X.Z.); yjhgxd@stumail.nwu.edu.cn (J.Y.); sjjnwu@163.com (J.S.); jumm089@163.com (M.J.)

[2] School of Pharmacy, Xi'an Jiaotong University, Xi'an 710061, China; zhoutao196@mail.xjtu.edu.cn

[3] College of Biopharmaceutical and Food Engineering, Shangluo University, Shangluo 726000, China; yezi19820320@163.com

* Correspondence: gfzhao@nwu.edu.cn; Tel.: +86-029-8830-5264

Received: 18 July 2018; Accepted: 24 August 2018; Published: 28 August 2018

Abstract: Cucurbitaceae is the fourth most important economic plant family with creeping herbaceous species mainly distributed in tropical and subtropical regions. Here, we described and compared the complete chloroplast genome sequences of ten representative species from Cucurbitaceae. The lengths of the ten complete chloroplast genomes ranged from 155,293 bp (*C. sativus*) to 158,844 bp (*M. charantia*), and they shared the most common genomic features. 618 repeats of three categories and 813 microsatellites were found. Sequence divergence analysis showed that the coding and IR regions were highly conserved. Three protein-coding genes (*accD*, *clpP*, and *matK*) were under selection and their coding proteins often have functions in chloroplast protein synthesis, gene transcription, energy transformation, and plant development. An unconventional translation initiation codon of *psbL* gene was found and provided evidence for RNA editing. Applying BI and ML methods, phylogenetic analysis strongly supported the position of *Gomphogyne*, *Hemsleya*, and *Gynostemma* as the relatively original lineage in Cucurbitaceae. This study suggested that the complete chloroplast genome sequences were useful for phylogenetic studies. It would also determine potential molecular markers and candidate DNA barcodes for coming studies and enrich the valuable complete chloroplast genome resources of Cucurbitaceae.

Keywords: Cucurbitaceae; chloroplast genome; structural comparison; selective pressures; RNA editing; phylogeny

1. Introduction

As the fourth most economically important plant family, Cucurbitaceae consists of 115 proposed genera with approximately 960 species distributed in tropical and subtropical areas [1]. The vast majority of these plants are annual vines and woody lianas, and only a small proportion are shrubs and trees [2]. Cultivars developed by breeders, especially melon (*Cucumis melo*), watermelon (*Citrullus lanatus*), bottle gourd (*Lagenaria siceraria*), pumpkin (*Cucurbita pepo*), and cucumber (*Cucumis sativus*), are the basis for industries [3]. Their fruits are not only edible, but also used by humans mostly as durable containers, fishnet floats, and musical instruments [4]. The commercial use of derivatives from medicinal species is increasing rapidly. For instance, flavonoids and saponins contained in *Gynostemma pentaphyllum* [5] have radical scavenging and antiproliferative properties [6], and cucurbitane-type compounds extracted from *Hemsleya amabilis* and *Hemsleya carnosiflora* exert anti-inflammatory functions in bronchitis and tuberculosis treatments [7–9]. These plants are also used as traditional Chinese medicinal herbs because of their anticancer effect [6,10]. Therefore, over the last decades, large amounts of research have paid much attention to the improvement of cultivated varieties and the development of medicinal value. However, the most important basis of developing natural medicine is the wild species' identification, which is also very difficult. Take the genera *Gomphogyne*, *Hemsleya*, and *Gynostemma* as examples: they are all morphologically creeping and herbaceous with 3-11-foliolate leaves in above-ground plants. Although *Gomphogyne* ismonoecious, *Gynostemma* is dioecious, and *Hemsleya* has enlarged underground tubers, it has been rather problematic to define the classification of these species in the wild, especially without flowering or excavation [11].

In addition, the existing studies about Cucurbitaceae mainly focus on the history of domestication, origin, and dispersal [1,12–15]. Nevertheless, the phylogeny of Cucurbitaceae family has not yet been clearly solved. Due to the description and validation of new species, the number of interspecies and the attribution problem of some genera remain uncertain, such as with *Hemsleya* and *Gomphogyne.* Although some molecular-based phylogenic studies have been carried out on many Cucurbitaceae genera, *Hemsleya* and *Gomphogyne* were either not involved [16–18] or just participated in systematic surveys based on some specific fragments of DNA with a limited number of species within each genus [19,20]. Therefore, we prefer using the whole complete chloroplast genomes (CPGs) to resolve the phylogenic problem of the genera in the Cucurbitaceae family. Meanwhile, more DNA barcodes from genomic resources are in demand, for use in the identification of species among genera in the Cucurbitaceae family, and in further studies to reveal the genetic diversity, population structure, origin, and evolution of these species.

Comparatively speaking, genome-wide datasets have an edge over traditional DNA markers in providing information to effectively solve historically complex phylogenetic relationships [21–23]. The chloroplast genome (CPG) is a circular double-stranded DNA molecule which has maternal inheritance in the majority of plants [24,25]. It is smaller than the nuclear genome in size, and has a moderate rate of nucleotide evolution, but shows a difference in the rate of divergence between protein coding (CDS) and noncoding (CNS) regions [26]. Previous studies have demonstrated that most CPGs of angiosperms have a stable quadripartite structure: a pair of inverted repeats regions (IRa and IRb), one large single-copy region (LSC) and one small single-copy region (SSC) [24]. The common sizes of CPGs range from 120 kb to 160 kb usually caused by contractions and expansions of IR regions [27]. The comparative analyses of complete CPGs could contribute to understanding the complete CPG structure and evolution, identification of species and phylogenetic relationships [28].

In this study, comparative analyses were applied on complete CPGs of ten representative species in Cucurbitaceae to explore structural differentiation and molecular evolution of CPG sequences and increase the number of valuable complete CPG resources. The characterization of highly variable regions would contribute to developing candidate DNA barcodes for future studies. Microsatellites (SSRs) could be used as potential molecular polymorphic markers to reveal the genetic diversity and population structure of Cucurbitaceae. The identification of protein-coding genes under selection would play an important role in the analyses of adaptive evolution for plants in ecosystems. Furthermore, this study would reconstruct the intergeneric relationships and locate the phylogenetic position of genera *Gomphogyne* and *Hemsleya* in Cucurbitaceae.

2. Results

2.1. Genome Features

For three newly-obtained CPGs, the mean coverage of raw reads ranged from 1112.4 to 1392.7 (Table 1), and the lengths of consensus sequences were 157,334 bp (*G. cissiformis* var. *cissiformis*), 156,585 bp (*G. cissiformis* var. *villosa*) and 158,275 bp (*H. lijiangensis*). Each of them encoded 133 genes including 87 protein-coding genes, eight rRNA genes, 37 tRNA genes, and one pseudogene (Tables 1 and 2; Figure 1).

The comparative analyses of whole CPGs from ten species of Cucurbitaceae showed that the sizes of 10 CPGs ranged from 155,293 bp (*C. sativus*) to 158,844 bp (*M. charantia*), with an average CPG sequence length of 157,264 bp. All of the CPGs displayed a typical quadripartite structure: an LSC region (ranged from 86,642 bp to 88,374 bp) and an SSC region (ranged from 17,897 bp to 18,653 bp) which were separated by two IR regions (ranged from 25,193 bp to 26,242 bp; Table 1, Figure 2). The LSC region and IR region had a significant correlational relationship with the overall genome size, and each of the structural regions of the CPGs were not correlated with each other (Figure 2). A comparison of CPG sequences among ten species showed that there was no dramatic difference in compared features. The GC content percentage of *C. lanatus* (37.2%) was more than any of the other genomes (36.7–37.1%), while *M. charantia* had the lowest GC content (36.7%). For four structural regions, the GC content of IR region (42.7–43.1%) was clearly higher than that of the LSC (34.3–34.9%) region and SSC (30.6–31.8%) region for each CPG (Table 1). The CPGs encoded 122 to 135 functional genes including some pseudogenes, i.e., the *infA* gene in *G. cissiformis* var. *cissiformis*, *G. cissiformis* var. *villosa*, *H. lijiangensis*, and *G. pentaphyllum*, the *rps16* gene in *C. sativus*, and the *ycf1* gene in *M. charantia* and *L. siceraria* (Tables 1 and 2).

Table 1. Genome features of the chloroplast genomes of ten Cucurbitaceae species.

Species	* *G. cissiformis* var. *cissiformis*	* *G. cissiformis* var. *villosa*	* *H. lijiangensis*	*G. pentaphyllum*	*C. lanatus*	*C. grandis*	*C. sativus*	*C. moschata*	*M. charantia*	*L. siceraria*
Locations	24.20° N, 99.50° E	24.20° N, 99.50° E	27.17° N, 100.06° E	/	/	/	/	/	/	/
Assembly reads	1,274,004	1,505,442	1,483,091	/	/	/	/	/	/	/
Mean coverage	1112.4	1358.4	1392.7	/	/	/	/	/	/	/
Size (bp)	157,334	156,585	158,275	157,576	156,906	157,035	155,293	157,644	158,844	157,145
LSC (bp)	87,239	86,642	87,362	86,757	86,845	86,749	86,688	88,343	88,374	86,843
SSC (bp)	18,014	18,029	18,429	18,653	17,897	18,004	18,223	18,156	18,010	18,008
IRs (bp)	26,041	25,957	26,242	26,083	26,082	26,141	25,193	25,573	26,228	26,147
Number of total genes	133	133	133	133	122	132	132	135	130	130
Number of protein-coding genes	87	87	87	87	85	85	85	85	85	86
Number of tRNA genes	37	37	37	37	29	39	38	42	38	37
Number of rRNA genes	8	8	8	8	8	8	8	8	8	8
Pseudogene	*infA*	*infA*	*infA*	*infA*	/	/	*rps16*	/	*ycf1*	*ycf1*
Overall GC content (%)	37	37	37	37	37.2	37.1	37.1	37.1	36.7	37.1
GC content in LSC (%)	34.8	34.8	34.8	34.8	34.9	34.8	34.8	34.9	34.3	34.9
GC content in SSC (%)	31.1	31	31	30.6	31.5	31.3	31.8	31.5	30.7	31.4
GC content in IR (%)	42.8	42.7	42.8	42.8	42.8	42.8	42.8	43.1	42.8	42.8
GenBank number	MH256801	MF784515	MG733988	KX852298	KY014105	KX147312	AJ970307	MF991116	MG022622	MG022623

* Three newly obtained chloroplast genomes.

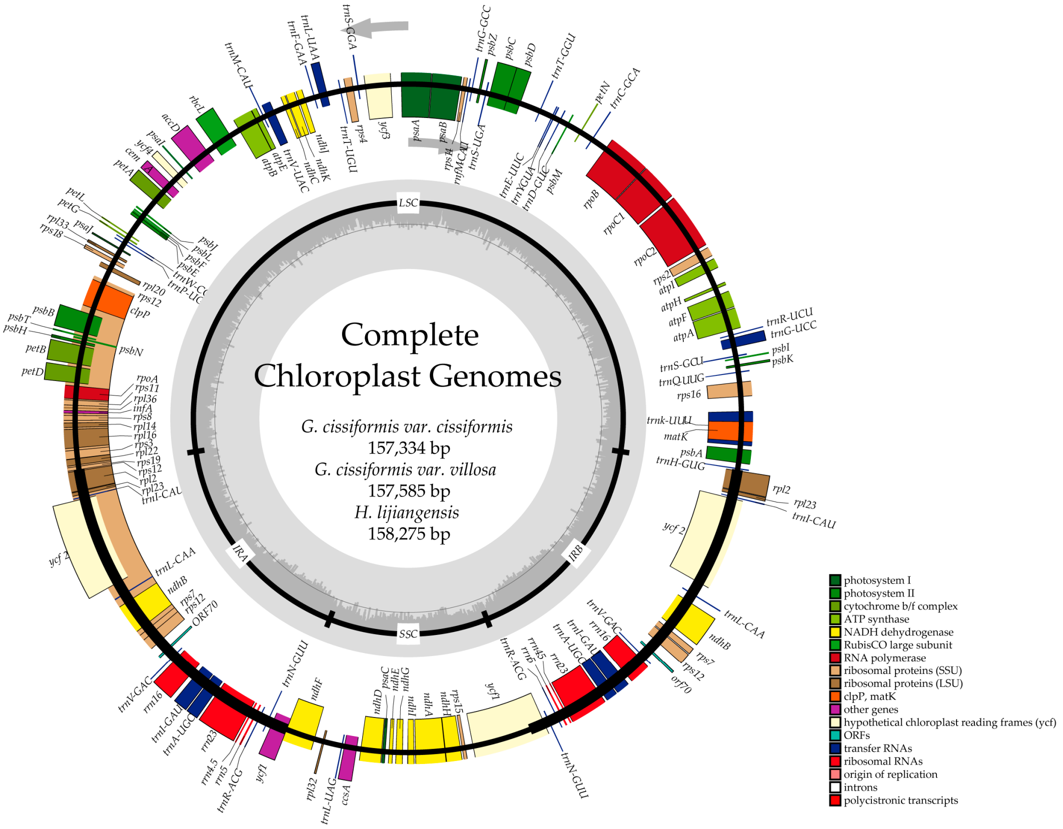

Figure 1. Gene maps of chloroplast genomes of Cucurbitaceae. Genes on the inside of the large circle are transcribed clockwise and those on the outside are transcribed counterclockwise. The genes are color-coded based on their functions. Dashed area represents the GC composition of the chloroplast genome.

Table 2. List of genes in the chloroplast genome of three newlysequenced species.

Category	Gene Group	Gene Name
Self-replication	Ribosomal protein (small subunit) (14)	*rps2 rps3 rps4 rps7* (×2) *rps8 rps11* * *rps12* (×2) *rps14 rps15* * *rps16 rps18 rps19*
	Ribosomal protein (large subunit) (11)	* *rpl2* (×2) *rpl14* * *rpl16 rpl20 rpl22 rpl23* (×2) *rpl32 rpl33 rpl36*
	RNA polymerase (4)	*rpoA rpoB* * *rpoC1 rpoC2*
	Transfer RNAs (37)	* *trnA-UGC* (×2) *trnC-GCA trnD-GUC trnE-UUC trnF-GAA trnfM-CAU* * *trnG-UCC trnG-GCC trnH-GUG trnI-CAU*(×2) * *trnI-GAU* (×2) * *trnK-UUU trnL-CAA*(×2) *trnL-UAG* * *trnL-UAA trnM-CAU trnN-GUU*(×2) *trnP-UGG trnQ-UUG trnR-ACG*(×2) *trnR-UCU trnS-GCU trnS-GGA trnS-UGA trnT-GGU trnT-UGU trnV-GAC*(×2) * *trnV-UAC trnW-CCA trnY-GUA*
	Ribosomal RNAs (8)	*rrn4.5*(×2) *rrn5*(×2) *rrn16*(×2) *rrn23*(×2)
Photosynthesis	Photosystem I (5)	*psaA psaB psaC psaI psaJ*
	Photosystem II (15)	*psbA psbB psbC psbD psbE psbF psbH psbI psbJ psbK psbL psbM psbN psbT psbZ*
	Cytochrome b/f complex (6)	*petA* * *petB* * *petD petG petL petN*
	ATP synthase (6)	*atpA atpB atpE* * *atpF atpH atpI*

Table 2. *Cont.*

Category	Gene Group	Gene Name
	NADH dehydrogenase (12)	* *ndhA* * *ndhB* (×2) *ndhC ndhD ndhE ndhF ndhG ndhH ndhI ndhJ ndhK*
	Rubisco large subunit (1)	*rbcL*
Other genes	Maturase (1)	*matK*
	membrane protein (1)	*cemA*
	Acetyl-CoA carboxylase gene (1)	*accD*
	ATP-dependent protease subunit (1)	*clpP*
	c-type Cytochrome biogenesis (1)	*ccsA*
	Assembly/stability of photosystem I (2)	*ycf3 ycf4*
	Conserved reading frames (ycfs) (4)	*ycf1*(×2) *ycf2*(×2)
	hypothetical chloroplast protein (2)	*orf70*(×2)
Pseudogene	Translation-related gene (1)	*infA*

* Gene with intron(s).

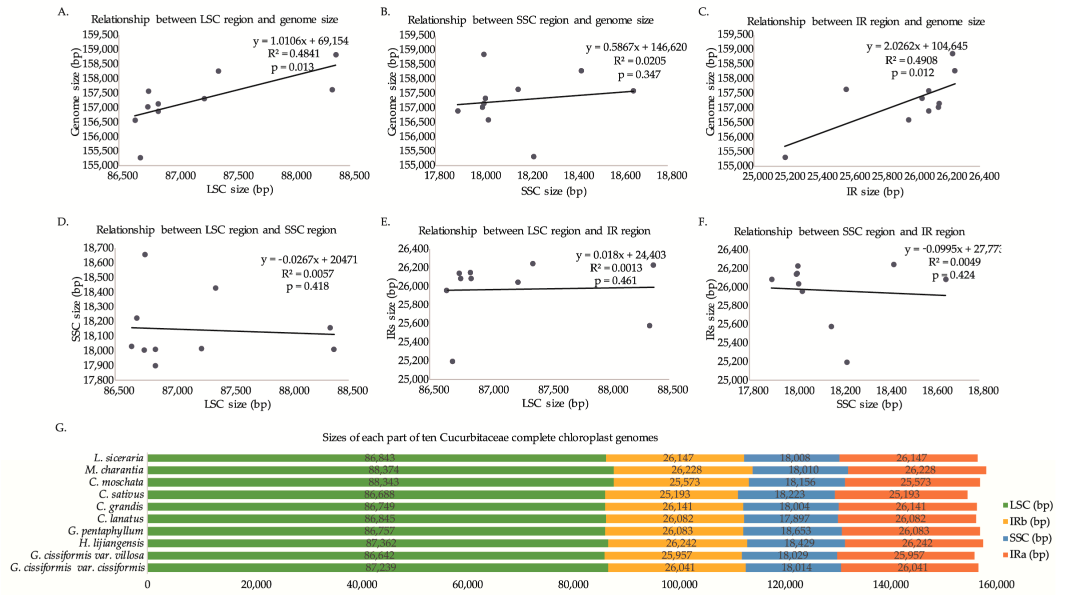

Figure 2. Relationships between complete chloroplast genome sizes and LSC, SSC and IR regions lengths, respectively. (**A**–**F**) Correlational relationships among each region; (**G**) sizes of each part of ten Cucurbitaceae complete chloroplast genomes.

2.2. IR/SC Boundary, Genome Rearrangement and Sequence Divergence

The IR/SC boundary areas of 10 CPGs of Cucurbitaceae species and two outgroups were compared (Figure 3). The gene content and order were observed to have some differences, for example, gene *ycf1* and gene *rpl2* were lost in two LSC borders of *M. charantia* and *L. siceraria*; gene *orf224* existed in the IRb border of *C. sativus* and *C. grandis*; as well as the location of gene *rps19* was diversified in all of the examined species (Figure 3). The expansions and contractions of IR region were discovered. Taking *M. charantia* as an example, gene *rps19* that was located in the LSC region was 207 bp away from the LSC/IRb boundary, while this distance was 0–6 bp in some other species, and gene *rpl12* was located in the IRb region, straddling the LSC/IRb border. Gene *ycf1*, located in IRb region, had 131 bp beyond the IRb/SSC boundary, and the comparable region was 10–12 bp long in some other species. This indicated that the relative position of the LSC/IRb boundary had moved backwards, and the IRb/SSC boundary, forwards. Correspondingly, the *ycf1* gene, located in SSC region, had just 29 bp across the SSC/IRa boundary, while this similar region was 971–1186 bp in most of the other species. Both of these phenomena demonstrated a contraction of two IR regions in the complete CPGs.

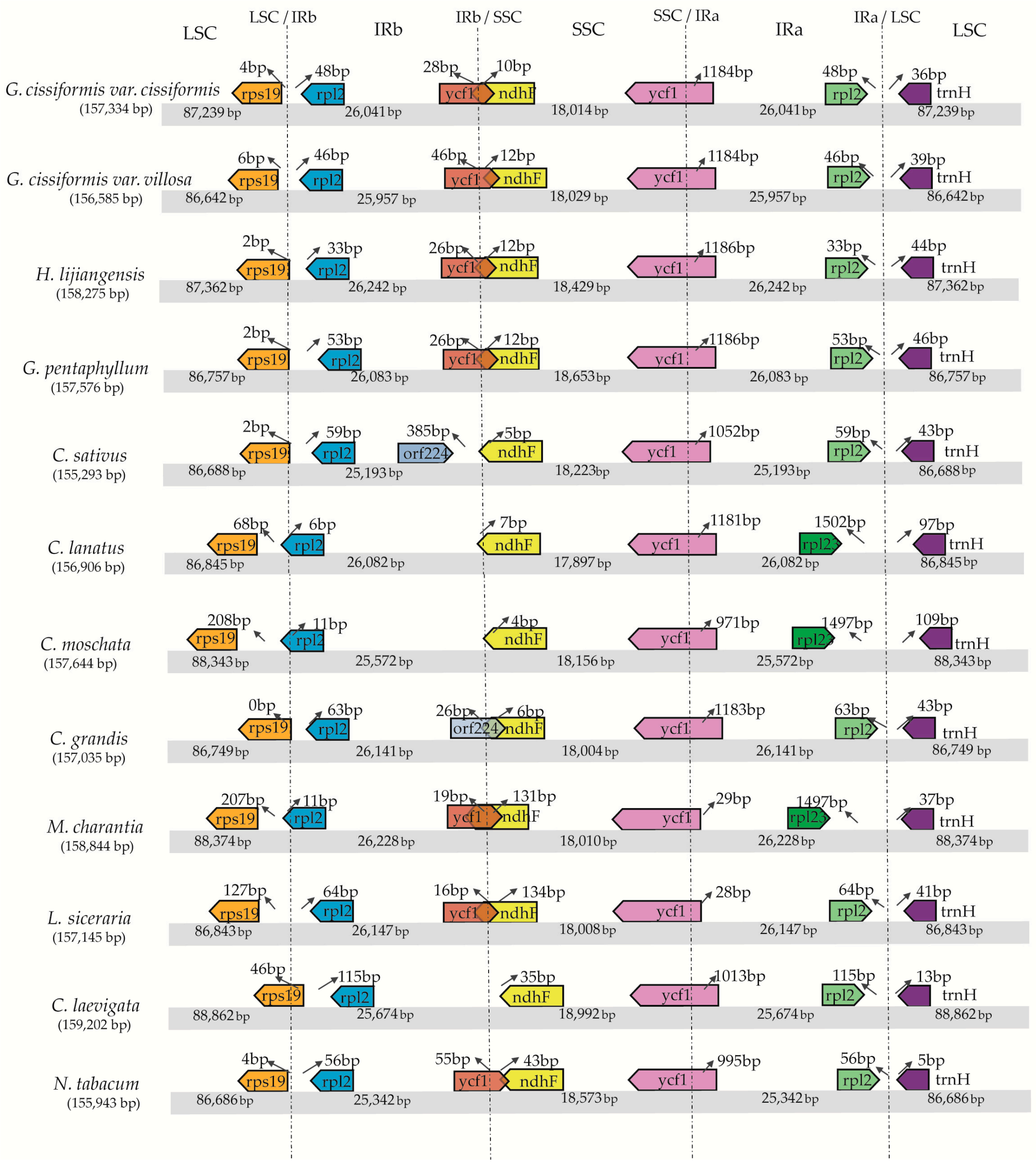
LSC
LSC / IRb
IRb
IRb / SSC
SSC
SSC / IRa
IRa
IRa / LSC
LSC
G. cissiformis var. cissiformis (157,334 bp)
4bp 48bp 28bp 10bp 1184bp 48bp 36bp
rps19 rpl2 ycf1 ndhF ycf1 rpl2 trnH
87,239 bp 26,041 bp 18,014 bp 26,041 bp 87,239 bp
G. cissiformis var. villosa (156,585 bp)
6bp 46bp 46bp 12bp 1184bp 46bp 39bp
rps19 rpl2 ycf1 ndhF ycf1 rpl2 trnH
86,642 bp 25,957 bp 18,029 bp 25,957 bp 86,642 bp
H. lijiangensis (158,275 bp)
2bp 33bp 26bp 12bp 1186bp 33bp 44bp
rps19 rpl2 ycf1 ndhF ycf1 rpl2 trnH
87,362 bp 26,242 bp 18,429 bp 26,242 bp 87,362 bp
G. pentaphyllum (157,576 bp)
2bp 53bp 26bp 12bp 1186bp 53bp 46bp
rps19 rpl2 ycf1 ndhF ycf1 rpl2 trnH
86,757 bp 26,083 bp 18,653 bp 26,083 bp 86,757 bp
C. sativus (155,293 bp)
2bp 59bp 385bp 5bp 1052bp 59bp 43bp
rps19 rpl2 orf224 ndhF ycf1 rpl2 trnH
86,688 bp 25,193 bp 18,223 bp 25,193 bp 86,688 bp
C. lanatus (156,906 bp)
68bp 6bp 7bp 1181bp 1502bp 97bp
rps19 rpl2 ndhF ycf1 rpl2 trnH
86,845 bp 26,082 bp 17,897 bp 26,082 bp 86,845 bp
C. moschata (157,644 bp)
208bp 11bp 4bp 971bp 1497bp 109bp
rps19 rpl2 ndhF ycf1 rpl2 trnH
88,343 bp 25,572 bp 18,156 bp 25,572 bp 88,343 bp
C. grandis (157,035 bp)
0bp 63bp 26bp 6bp 1183bp 63bp 43bp
rps19 rpl2 orf224 ndhF ycf1 rpl2 trnH
86,749 bp 26,141 bp 18,004 bp 26,141 bp 86,749 bp
M. charantia (158,844 bp)
207bp 11bp 19bp 131bp 29bp 1497bp 37bp
rps19 rpl2 ycf1 ndhF ycf1 rpl2 trnH
88,374 bp 26,228 bp 18,010 bp 26,228 bp 88,374 bp
L. siceraria (157,145 bp)
127bp 64bp 16bp 134bp 28bp 64bp 41bp
rps19 rpl2 ycf1 ndhF ycf1 rpl2 trnH
86,843 bp 26,147 bp 18,008 bp 26,147 bp 86,843 bp
C. laevigata (159,202 bp)
46bp 115bp 35bp 1013bp 115bp 13bp
rps19 rpl2 ndhF ycf1 rpl2 trnH
88,862 bp 25,674 bp 18,992 bp 25,674 bp 88,862 bp
N. tabacum (155,943 bp)
4bp 56bp 55bp 43bp 995bp 56bp 5bp
rps19 rpl2 ycf1 ndhF ycf1 rpl2 trnH
86,686 bp 25,342 bp 18,573 bp 25,342 bp 86,686 bp

Figure 3. Comparison of the LSC, IR, and SSC border regions among the 10 Cucurbitaceae chloroplast genomes. Number above the gene features means the distance between the ends of genes and the borders sites. These features are not to scale.

The whole-genome alignment of the 10 CPGs showed that there were no rearrangement events in Cucurbitaceae (Figure S1). Using *G. cissiformis* var. *cissiformis* as the reference, the alignment of 11 Cucurbitales CPGs were performed to investigate the level of sequence divergence (Figure S2). The result showed a high sequence similarity within genus *Gomphogyne*, but great divergence among different genera and families. As expected, the SC and CNS regions exhibit more differences than IRs and CDS regions, respectively. Moreover, the percentage of variable features in coding and non-coding regions, and the percentage of indels in each variation, were calculated based on two patterns of sequence alignment: (1) two species within *Gomphogyne*; and (2) the represented species of nine genera within Cucurbitaceae. The results showed that the percentage of variable features of CNS ranged from 0 to 43.00% for two *Gomphogyne* species and from 0 to 89.25% for nine Cucurbitaceae genera, with an average level of 2.92% and 33.49% respectively. These percentages were much higher than those of the CDS: from 0 to 16.05% for two *Gomphogyne* species, and from 0 to 42.64% for nine Cucurbitaceae genera, with an average level of 0.41% and 7.49%, respectively (Figure 4, Table S3). This also indicated that CDS was much more conservative than CNS. Mostly, the variations were located in SC regions instead of IR regions. The results also suggested that the variations among nine genera were higher than those between species within a single genus, and most variations were caused by indels (Figure 4). In addition, the top- four highly-variable genes (*accD*, *rpl22*, *ycf1*, and *ycf1*) and top-four highly divergent intergenic regions (*trnR (UCU)-atpA*, *trnL (UAA)-trnF (GAA)*, *rpl32-trnL (UAG)*, and *ndhA intron*) were confirmed (Table S3) and primers for these regions were shown in Supplementary Table S4. These regions could be used as candidate DNA fragments for further studies related to genetics, phylogeny, and species identification.

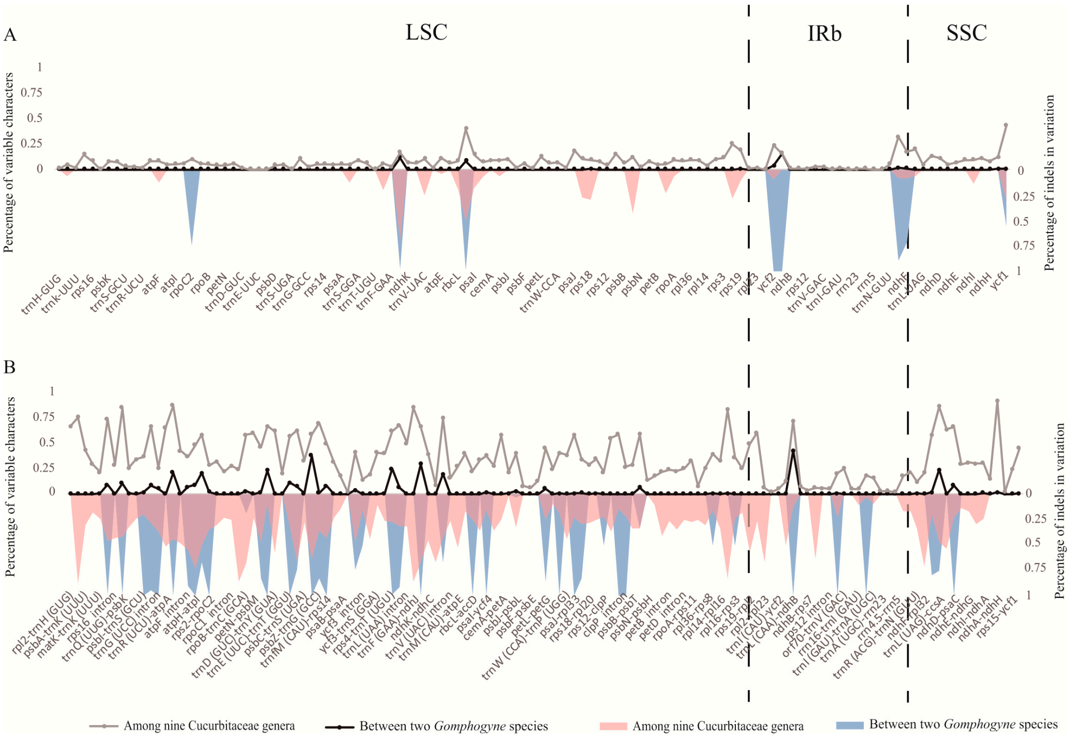

Figure 4. Percentage of variable characters in aligned Cucurbitaceae chloroplast genomes. (**A**) Coding region (CDS) and (**B**) Noncoding region (CNS). These regions are oriented according to their locations in the chloroplast genome.

2.3. Repeat Analysis and Microsatellites (SSR)

Three categories of repeats (tandem, dispersed, and palindromic repeats) were identified in the 11 Cucurbitales CPGs (Figure 5A, Tables S5, S6, and S8). A total of 618 repeats were identified for these species including 163 tandem repeats, 247 dispersed repeats, and 208 palindromic repeats, indicating the highest percentage (40%) of dispersed repeats (Figure 5B). Among different species, the number of repeats for *G. cissiformis* var. *cissiformis* (76) and *C. grandis* (36) were the most and the least, respectively (Figure 5A). Additionally, 813 SSRs were found, of which, the number of mono-, di-, tri-, tetra-, penta-, and hexanucleotide repeats were 552, 121, 37, 79, 18, and 6, respectively. It was shown that the mononucleotide repeats were most common, accounting for 68% of all, while the dinucleotides repeats accounted for 15%, and another polynucleotide SSRs occurred at less frequently (Figure 5D, Tables S7 and S8). From the perspective of species, *C. laevigata* (125) had the most SSRs and *G. cissiformis* var. *cissiformis* (58) had the least (Figure 5C).

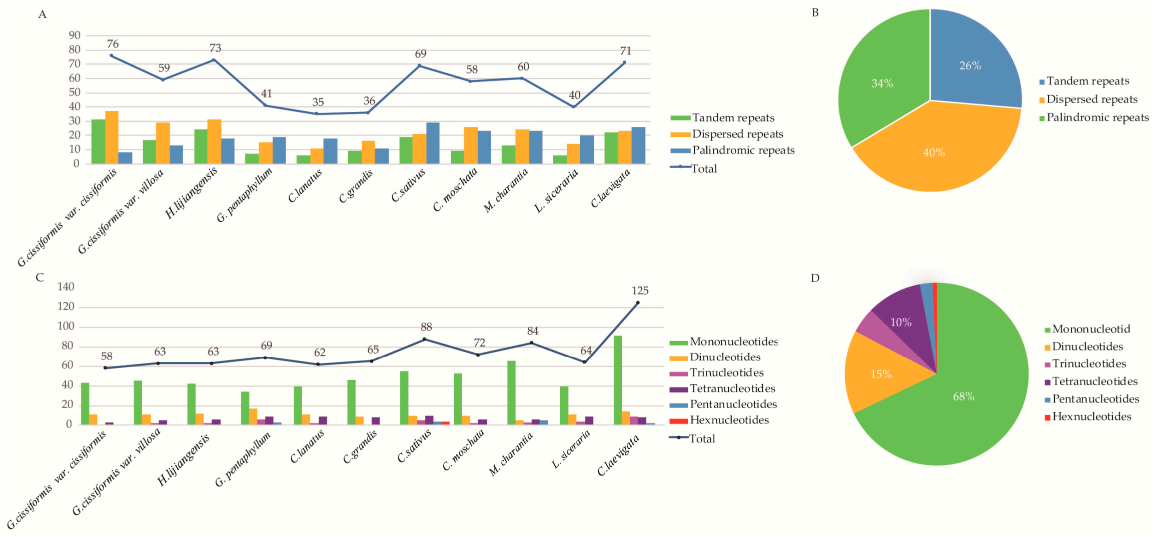

Figure 5. The type and presence of repeated sequences and simple sequence repeats (SSR) in the chloroplast genomes of eleven Cucurbitales species. (**A**) Number of three-types repeats; (**B**)Percentage of three repeat types; (**C**) Number of SSRs and their types; (**D**) Percentage of SSR types.

2.4. Selective Pressures Events

The non-synonymous (K_A) and synonymous (K_S) substitution ratio (K_A/K_S) were calculated for 68 consensus protein-coding genes to estimate selective pressures. Although all of the K_A/K_S (ω) values were less than 1.0 in codeml, the K_A/K_S ratio of five genes (*clpP*, *atpE*, *psbL*, *accD*, and *matK*) were within the range of 0.5 to 1.0 indicating a relaxed selection. Among them, the likelihood ratio test (LRT) analysis showed several sites from three genes (*accD*, *clpP*, and *matK*), which were distributed in the LSC region (Table 3, Figure 6), were under selection. We located the consistent selective sites under the naive empirical Bayes (NEB) and the Bayes empirical Bayes (BEB) methods in the alignment of CPGs, and found these amino acid sites had a high level of variation, for example, in the 308 site of *accD* gene, the codon CGG could have the variables CAG, CTG, AAG, and GAA (Figure 6). Unfortunately, there was only one K_A/K_S (ω) value that was greater than 1.0 (gene *atpE*), but no significant *p*-value ($p < 0.05$, Table S9) was found using the KaKs-calculator.

Table 3. Parameter estimates and log-likelihood values for different models in selective pressure analysis.

Genes	Model	df	lnL/ω Value	LRTs	No. of Sites (BEB)	Consistent Sites
clpP	M0 (one ratio)	19	**ω = 0.96975**			
	M1 (neutral)	20	−1396.7593	M1 vs. M2:	2	12 S/N/L
	M2 (selection)	22	−1389.6667	14.1852 **		AGT/AAT/CTT
	M7 (beta)	20	−1396.7638	M7 vs. M8:	4	87 R/K/S
	M8 (beta&ω)	22	−1389.6672	14.1933 **		CGA/AAA/TCA
atpE	M0 (one ratio)	19	**ω = 0.70941**			
	M1 (neutral)	20	−787.914712	M1 vs. M2:	0	/
	M2 (selection)	22	−785.703588	4.42225		
	M7 (beta)	20	−787.942937	M7 vs. M8:	0	
	M8 (beta&ω)	22	−785.704169	4.47754		
psbL	M0 (one ratio)	19	**ω = 0.61775**			
	M1 (neutral)	20	−158.807141	M1 vs. M2:	0	/
	M2 (selection)	22	−158.807091	0.00010		
	M7 (beta)	20	−158.807137	M7 vs. M8:	0	
	M8 (beta&ω)	22	−158.80712	0.00003		

Table 3. *Cont.*

Genes	Model	df	lnL/ω Value	LRTs	No. of Sites (BEB)	Consistent Sites
accD	M0 (one ratio)	19	**ω = 0.53161**			
	M1 (neutral)	20	−3221.0459	M1 vs. M2:	1	308 R/Q/K/E/L
	M2 (selection)	22	−3214.0011	14.0896 **		CGG/CAG/CTG/AAG/GAA
	M7 (beta)	20	−3221.5205	M7 vs. M8:	4	
	M8 (beta&ω)	22	−3214.0768	14.8875 **		
matK	M0 (one ratio)	19	**ω = 0.52255**			
	M1 (neutral)	20	−3688.5152	M1 vs. M2:	1	337 T/I/A
	M2 (selection)	22	−3683.8284	9.3735 **		TCA/GGA/GCA
	M7 (beta)	20	−3689.0422	M7 vs. M8:	8	
	M8 (beta&ω)	22	−3683.9212	10.2421 **		

** $p < 0.01$; df: degree of freedom; the likelihood ratio tests, LRT = ∣df(M2/M8) − df(M1/M7)∣ × ∣lnL(M2/M8) − lnL(M1/M7)∣; No. of Sites: the number of selective sites under the Bayes empirical Bayes (BEB) model; Consistent sites: the sites appeared in both M1 vs. M2 and M7 vs. M8, showing the amino acids and their corresponding codons.

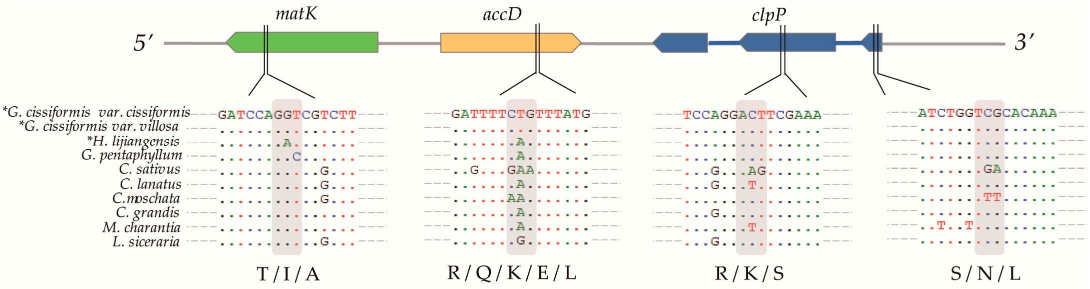

Figure 6. Alignment of selective sites of 10 Cucurbitaceae species. * marked three newly obtained CPGs.

2.5. Codon Usage Bias and Unconventional Initiation Codon

Codon usage of the protein-coding genes was analyzed in the CPGs of 10 Cucurbitaceae species. The number of encoded codons ranged from 25,922 (*C. sativus*) to 26,828 (*G. pentaphyllum*) (Table S10). Detailed codon analysis showed that the 10 Cucurbitaceae species had a similar codon constituent, and close RSCU (relative synonymous codon usage) values (Table S10). Leucine (Leu) and Cysteine (Cys) were the highest (10.60%) and lowest (1.20%) frequently used amino acids in these species, respectively (Figure S3A, Table S10). The results revealed that most of the amino acid codons have preferences with the exception of Met (Methionine—AUG) and Trp (Tryptophan—UGG). Three newly obtained CPGs had 31 biased codons with RSCU > 1, while other CPGs had 30 (Table S10) due to the difference in codon Ser (Serine—UCC), which were used more than 350 times in the three species and less than 336 times in other species (Table S10). It was illustrated that the genera *Gomphogyne* and *Hemsleya* preferred using Serine more than any other genera. The codons had lower representation rates for C or G at the third codon position, and the average GC content of the third codon base was 37.8%, with the range from 37.6% to 38.0% (Table S11). It turned out that the CPGs of Cucurbitaceae species had a strong bias toward A or T at the third codon position.

Interestingly, an unconventional initiation codon (Thr-TGC) of *psbL* gene was stumbled on when three new species sequences were annotated. This phenomenon was also found in *G. pentaphyllum* after a global alignment of ten Cucurbitaceae species. When the sequences including the start codon of the *psbL* gene were blasting with the transcriptome dataset of *G. pentaphyllum*, it was indeed found that the ACG start codon had been converted into an initiation codon, AUG (Figure 7). This could be explained by the occurrence of RNA editing phenomenon during the translation process, which reverted this change.

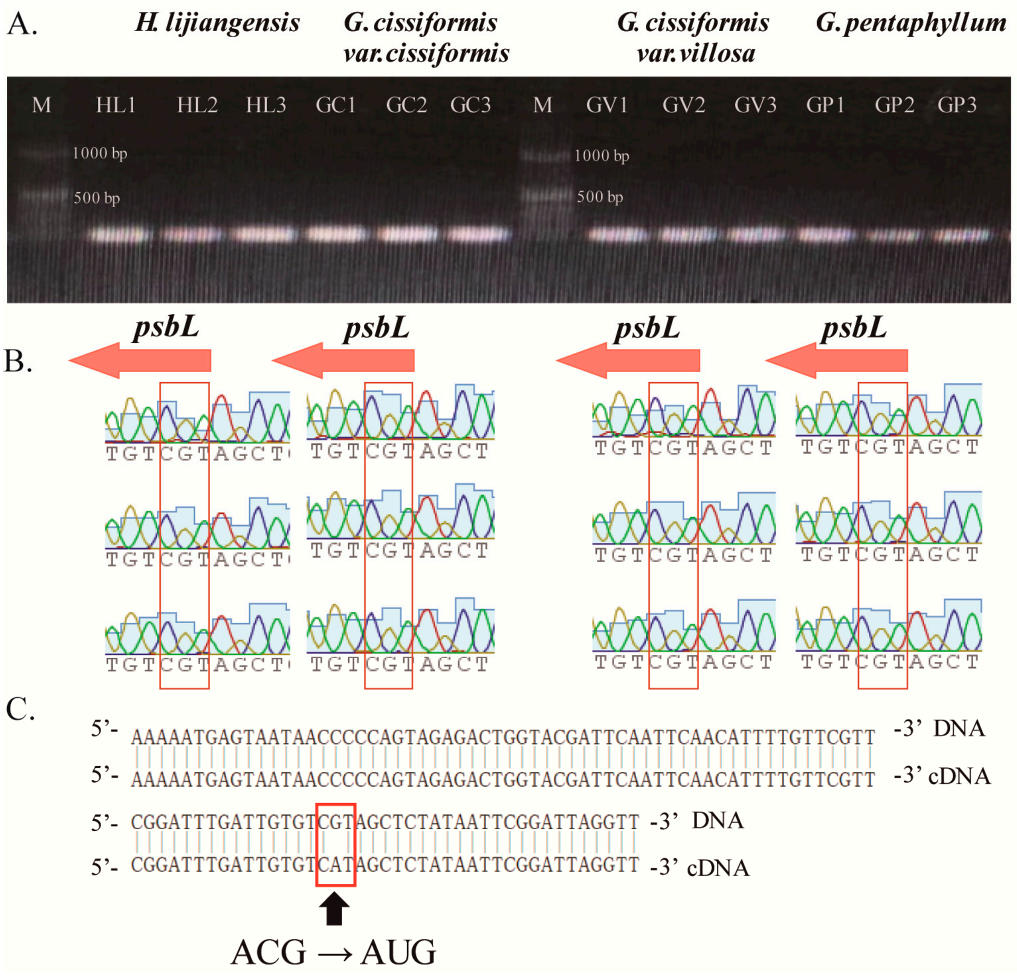
A.
H. lijiangensis
G. cissiformis var. cissiformis
G. cissiformis var. villosa
G. pentaphyllum
M HL1 HL2 HL3 GC1 GC2 GC3 M GV1 GV2 GV3 GP1 GP2 GP3
1000 bp
500 bp
1000 bp
500 bp
B.
psbL
psbL
psbL
psbL
TGTCGTAGCT
C.
5'- AAAAATGAGTAATAACCCCCAGTAGAGACTGGTACGATTCAATTCAACATTTTGTTCGTT -3' DNA
5'- AAAAATGAGTAATAACCCCCAGTAGAGACTGGTACGATTCAATTCAACATTTTGTTCGTT -3' cDNA
5'- CGGATTTGATTGTGTCGTAGCTCTATAATTCGGATTAGGTT -3' DNA
5'- CGGATTTGATTGTGTCATAGCTCTATAATTCGGATTAGGTT -3' cDNA
ACG → AUG

Figure 7. Result of unconventional initiation codon. (**A**) The PCR products electrophoresed in 1% agarose gel; (**B**) The PCR products sequences including start codon of gene *psbL*; (**C**)Blast results of sequence fragment of *G. pentaphyllum* including the start codon of *psbL* in the published transcriptome dataset.

2.6. Phylogenetic Analysis

All the ML and BI trees were reconstructed based on five datasets with the species of Cucurbitaceae released in NCBI. The best-fit models of ML and BI trees using the overall CPGs were GTR + I + G and TVM + I + G, respectively, and that for other datasets were displayed above the tree clade in Figure S4. It was shown that the phylogeny produced from the analyses of 27 complete CPG sequences was well-supported. All nodes of the phylogenetic tree were strongly supported by the 1.00 Bayesian posterior probabilities in BI analysis and 83–100% bootstrap values in ML analysis (Figure 8). It was shown that plants of Cucurbitaceae were clustered into one clade. Genera *Hemsleya* and *Gomphogyne*, constituted the earliest diverging lineage in this group, holding the closest relationship with genus *Gynostemma* and this clade was identified as a sister to all other species. Although there were some variations embodied in the phylogenetic positions of *C. rehmii* in *Citrullus* and *G. pentagynum* in *Gynostemma*, the phylogenetic relationships of any other species were concordant among the genera of Cucurbitales (Figure S4).

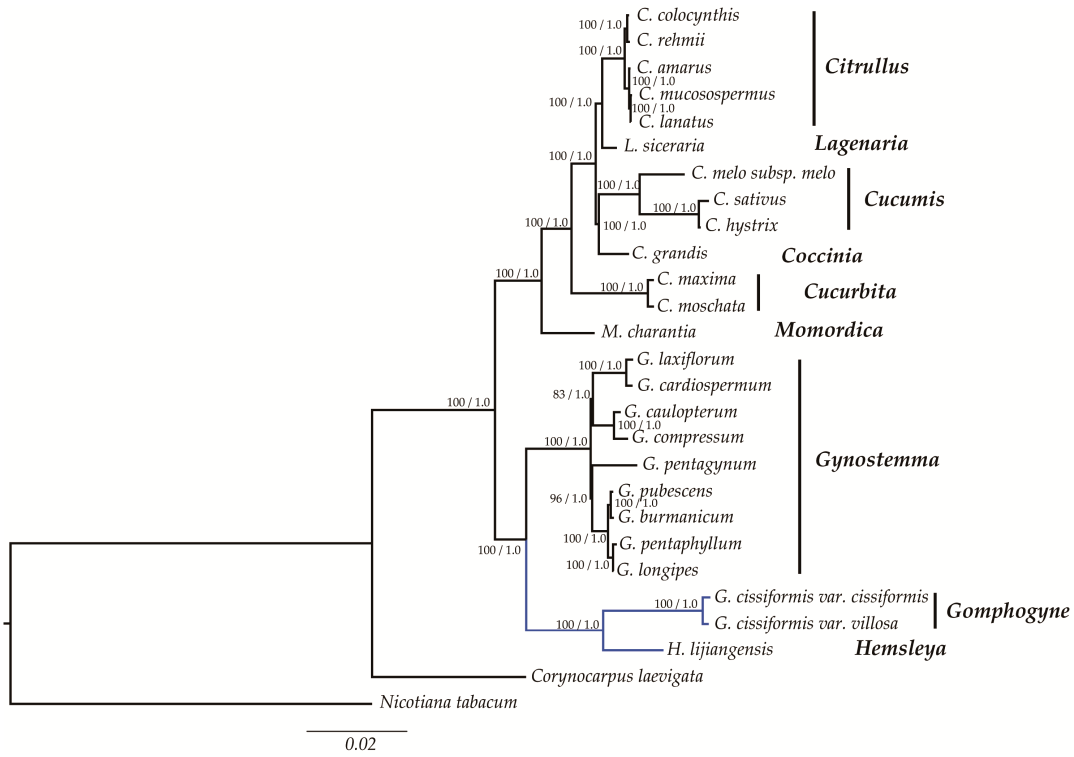

Figure 8. Phylogenetic relationship of the 27 species inferred from ML and BI analyses based on the complete cp genome sequences. The bootstrap values of ML analyses and Bayesian posterior probabilities are shown beside the clades. *Corynocarpus laevigata*, and *Nicotiana abacum* were used as outgroups. The clades in blue color showed the three newly sequenced species in our study.

3. Discussion

3.1. Evolution and Variation of Chloroplast Sequences

In angiosperms, most of the CPGs have evolved rapidly [29] and have some structural changes, such as gene rearrangements [30], gene loss-and-gain [31] and gene inversion [32], but no rearrangements events were found in any of our species after global alignment with the published CPGs, even if they contained a large number of large repeat sequences, which may be a reaction to the rearrangement of CPGs and sequence divergence in some other studies [33,34]. All ten CPGs that we studied displayed a typical quadripartite structure, with two SCs and two IRs arranged at regular intervals, and a highly conserved in genome structure and gene order., The pseudogene was initially thought to have lost the ability of protein coding [35] but was, instead, an evolutionary relic of the functional component [36]. In the present study, genes *ycf1* and *rpl2* were both lost in our species, while the *ycf1* gene was found to be a pseudogene in *M. charantia* and *L. siceraria*, but existed in *G. cissiformis* var. *cissiformis*, *G. cissiformis* var. *villosa*, *H. lijiangensis*, and *G. pentaphyllum*. In the complete CPGs of these aforementioned four species, plus *Syzygium cumini* and *Ananas comosus* [37,38], gene *infA* was also regarded as a pseudogene, and it was also found to be lost in the CPGs of *Alstroemeria aurea* and *Arabidopsis thaliana* [39,40].

The IR regions are highly conserved, and they are important in the stabilization of the CPG structure [41]. This is a common evolutionary phenomenon in plants and mainly reflected in the variation of CPGs in length [42,43]. Our results from the comparison of IR/SC boundary areas among species also suggested expansions and contractions of the IR region. As expected, both mVISTA and sequence divergence analysis indicated that CDS and IRs were more conserved than CNS and SCs. The sequence divergence also revealed many significant differences among the CPGs of the family, but a low level of differentiation between species within the genus *Gomphogyne*. When constructing phylogenetic trees with the sequences of four highly variable genes and four highly divergent intergenic regions of each CPG, the results were derived from the phylogenetic analyses based on the entire CPGs (Figure S4). These results indicated that the highly variable regions could be used in the phylogenetic analyses of Cucurbitaceae. Further work is still necessary to determine whether these highly variable regions could serve as candidate DNA barcodes to identify species. SSRs, which are also called microsatellites, can be used to analyze the genetic diversity, population structure, and phylogeography based on polymorphisms [26,44]. Thus, the SSR sequences we identified could contribute to molecular and evolutionary ecological knowledge, which warrants further research at the population level.

3.2. Selective Genes and RNA Editing

Analysis of the adaptive evolution of genes has an important reference value in examining the change of gene structure and functional mutations [45]. The percentage of nonsynonymous (K_A) versus synonymous (K_S) nucleotide substitutions (denoted by K_A/K_S, or ω value) is usually used to evaluate the rate of gene divergence, and determine whether positive, purifying, or neutral selection has been in operation [46]. The K_A/K_S ratio may reveal the constraints of natural selection on organisms, and the estimation of these mutations contribute greatly to understanding the dynamics of molecular evolution [47]. If $\omega > 1.0$, the corresponding genes experience positive selection, while $0.5 < \omega < 1.0$, and $\omega < 0.5$ indicate relaxed selection and purifying selection, respectively [48]. Among our calculations, there were five genes under relaxed selection ($0.5 < \omega < 1.0$, Table 3), and several selective sites were found in three (*accD*, *clpP*, and *matK*) of the genes.

It is well established that acetyl-CoA carboxylase (ACCase, EC 6.4.1.2) catalyzes the formation of malonyl-CoA from acetyl-CoA, and it is considered to be the regulatory enzyme of fatty acid synthesis [49,50]. The *accD* gene exactly encodes the β-carboxyl transferase subunit of acetyl-CoA carboxylase [51,52]. It is an essential gene required for leaf development [50], and has great effects on leaf longevity and seed yield [53]. However, this gene has been lost, or defined as a pseudogene, in some species of Primulaceae, Acoraceae, and Poales [22,54]. The *clpP* gene encodes the ATP-dependent clp protease proteolytic subunit [55]. This protein is an essential component to form the protein complex of clp protease (endopeptidase clp) which is active and probably involved in the turnover of chloroplast proteins [56]. It was reported that the loss of *clpP* gene product (the clpP protease subunit) would lead to ablation of the shoot system of tobacco plants, suggesting that clpP-mediated protein degradation is essential for shoot development [57,58]. The *matK* (maturase K) gene is a plant chloroplast gene [59] which is located within the intron of the *trnK* gene (Figure 1). The protein it encodes is an intron maturase which is involved in the cutting and splicing of Group II RNA transcriptional introns [60,61]. The *matK* retains only a well-conserved domain X, and remnants of a reverse transcriptase domain [61]. Usually, the *matK* gene sequence is effectively used as a DNA barcoding fragment for angiosperms, in studies of plant systematics [62–64].

In summary, the coding proteins of these selective genes were all enzymes functioning in chloroplast protein synthesis, gene transcription, energy transformation, and plant development. The majority of wild species in Cucurbitaceae are creeping herbs, mainly distributed in moist mountains, forests, thickets, and streamside, and they may have some mechanisms for adapting to complex living conditions. Therefore, the species may have produced some corresponding differentiation in morphology during the long process of evolution. Consequently, we inferred that the chloroplast functional genes which were under selection might play key roles during the adaptation and development of the Cucurbitaceae species to terrestrial ecosystems.

Codon usage bias plays an important role in the evolution of CPG. The main factor that contributed to biased codon usage is the GC content, which is also important during the evolution of genomic structure, such as stability of replication, transcription, and translation [65,66]. The observed GC content level indicated that the CPGs in Cucurbitaceae were GC-lacking, and that there was a strong bias towards A/T at the third codon position, consistent with the existing CPG research [67–70].

The presence of translation-preferred codons might be the result of both mutation preference and natural selection during the CPG evolutionary process [71].

The genetic information in land plant chloroplast DNA is sometimes altered at the transcript level by a process known as RNA editing [72]. This process of the post-transcriptional modification of precursor RNAs to alter their nucleotide sequences [73]. It sometimes occurs through the insertion and deletion of nucleotides, or specific nucleotide substitution (mostly C to U conversion) [72]. Since the first evidence of RNA editing was found in chloroplast in the rpl2 transcript of maize [74], it has been hunted out and systematically studied in the protein-coding transcripts from many major lineages of land plants [75], such as *Arabidopsis thaliana* [76], *N. tabacum* [49], *Zea mays* [77], *Oryza sativa* [78], *Cucumis melo*, and *Cucurbita maxima* [79]. Most of the studies suggested that RNA editing occasionally created start or stop codons which shorten the size of translation products [72,80,81], even producing a new gene in one striking case [80]. Our results revealed an unconventional initiation codon in the *psbL* gene, having a function in producing the PSII-L protein in the photosystem II (PSII) complex [82], caused by nucleotide substitution which was generated by RNA editing on the second position of start codon (ACG to AUG). This phenomenon was also found in bell pepper [83], tobacco [84,85], spinach [73], and *Ampelopsis brevipedunculata* [86]. RNA editing is very common in plant chloroplast genomes. It can modify mutations, change reading frames, and regulate the expression of chloroplast genes [87], acting as a correction mechanism in the chloroplast of plants.

3.3. Phylogeny of Cucurbitaceae

The number of studies using complete CPG sequences for assessing phylogenetic relationships among angiosperms has been increasing rapidly [21,88–90]. Our phylogenetic trees (both ML and BI trees) indicated a clear relationship of the genera in Cucurbitaceae with high bootstrap values. The phylogenetic trees demonstrated that the genera *Gomphogyne*, *Hemsleya*, and *Gynostemma* constituted the earliest diverging lineage in Cucurbitaceae. This was consistent with the proposal that these three genera were relatively original genera belonging to the family Cucurbitaceae based on morphology [91]. All in all, our results suggested that the CPG data can effectively resolve the phylogenetic relationships of these genera in Cucurbitaceae. In fact, for this large family, our study was just a drop in the bucket. Some studies pointed out that the lack of samples might also affect the results of the phylogenetic analysis [22]. Unfortunately, our study could not roundly figure out the relationships among genera due to the limited sample size. More species from more genera should be included in the future. Furthermore, our phylogenetic study was based solely on chloroplast DNA. In order to comprehensively understand of the systematic evolution of Cucurbitaceae, nuclear DNA analyses are required to investigate the effect of gene introgression and hybridization on phylogeny. Our phylogenetic studies provided a valuable resource that should contribute to the future taxonomy, phylogeny, and evolutionary history studies of the Cucurbitaceae family.

4. Materials and Methods

4.1. Plant Materials and DNA Extraction

Healthy leaves of three species (*G. cissiformis* var. *cissiformis*, *G. cissiformis* var. *villosa*, and *H. lijiangensis*) were collected from adult plants in Yunnan province, China (Table 1). Voucher specimens were deposited in the Evolutionary Botany Laboratory of Northwest University (Shaanxi, China). Total genomic DNA were extracted from silica-dried leaf materials with simplified CTAB protocol [92]. Data from seven complete CPGs (Table 1) [26,93–95] were recovered from the National Center of Biotechnology Information (NCBI) in order to conduct the follow-up analyses.

4.2. Illumina Sequencing, Assembly, and Annotation

Illumina raw reads were collected using an Illumina Hiseq 2500 platform. The quality-trim with all of the raw reads was performed using CLC Genomics Workbench v7.5 (CLC bio, Aarhus, Denmark) with the default parameter set. The programs MITObim v1.7 (University of Oslo, Oslo, Norway) [96] and MIRA v4.0.2 (DKFZ, Heidelberg, Germany) [97] were used to perform the reference-guided assembly twice, to reconstruct the CPGs with published *G. pentaphyllum* (KX852298) and *C. melo* (JF412791) as references, respectively. A few gaps, dubious bases, and low-coverage regions in the assembled CPGs were corrected by Sanger sequencing, whereby pairs of primers were designed (Table S1) using Primer 3 version 4.0.0 (Whitehead Institute for Biomedical Research, Massachusetts, USA) [98]. The software DOGMA, Dual Organellar Genome Annotator (University of Texas at Austin, Austin, TX, USA) [99], was used to annotate the complete CPGs, and corrected by comparing with the complete CPGs of the references mentioned above using GENEIOUS R8 (Biomatters Ltd., Auckland, New Zealand). The circular CPG maps were drawn using online software OGDRAW (http://ogdraw.mpimp-golm.mpg.de) (Max planck Institute of Molecular Plant Physiology, Potsdam, Germany). All of the newly generated complete CPG sequences were submitted to GenBank (Table 1).

4.3. Comparison of Complete Chloroplast Genomes

The mVISTA (The Regents of the University of California, Oakland, CA, USA) [100] software was employed to discover the interspecific variation among the complete CPG sequences of eleven species (ten Cucurbitaceae species and *Corynocarpus laevigata*, Corynocarpaceae, HQ207704), and the alignments with annotations were visualized using *G. cissiformis* var. *cissiformis* as reference. In order to analyze the expansions and contractions, as well as the variation in junction regions among ten Cucurbitaceae species, the IR region borders and gene rearrangements were surveyed by the plug-in program, Mauve, in GENEIOUS R8. To analyze the bivariate correlational relationship between the overall CPG sizes and each of the structural regions of CPGs, i.e., LSC region, SSC region and IR region, we used IBM SPSS Statistics v21.0 (SPSS Inc., Chicago, IL, USA) with Pearson's one-tail test, and the significant value was $p < 0.05$.

4.4. Sequence Divergence

The multiple alignments of the CPGs were carried out using MAFFT version 7.017 (Osaka University, Suita, Japan) [101]. DnaSP v5.0 (Universitat de Barcelona, Barcelona, Spain) [102] was used to compute the variable sites across the complete CPGs, LSC, SSC, and IR regions of all the species. To investigate the sequence divergence patterns, MEGA 5.0 (Tokyo Metropolitan University, Tokyo, Japan) [103] was employed for statistical analysis of the variations of CPGs and percentage of indels among each region. The percentage of variable characters for each coding and noncoding region were calculated based on the method of Zhang [104].

4.5. Microsatellites and Repeated Sequences

Microsatellites (SSRs) and three categories of repeated sequences were detected in all eleven Cucurbitaceous species. The software, MISA (Institute of Plant Genetics and Crop Plant Research, Gatersleben, Germany)) [105] was utilized to seek the microsatellites (SSRs) with thresholds of 10, 5, 4, 3, 3, and 3, for mono-, di-, tri-, tetra-, penta-, and hexa-nucleotides, respectively. The online program, Tandem Repeats Finder (http://tandem.bu.edu/trf/trf.html) (Mount Sinai School of Medicine, New York, NY, USA) [106], was used to find the tandem repeat sequences, which were at least 10 bp in length. The alignment parameters for match, mismatch, and indels were set to be 2, 7, and 7, respectively. To search out the size and location of dispersed and palindromic repeats, the online program REPuter (https://bibiserv2.cebitec.uni-bielefeld.de/reputer) (University of Bielefeld, Bielefeld, Germany) [107] was performed with parameters of 30 bp minimal repeat size, and the similarity percentage of two repeat copies was set to at least 90%.

4.6. Selective Pressure Analysis

Selective pressures were analyzed for consensus protein-coding genes among ten Cucurbitaceae species. PAML with codeml program (University College London, London, UK) [108] was performed to calculate the nonsynonymous (K_A) and synonymous (K_S) substitution ratio. In order to estimate the ω value ($\omega = K_A/K_S$) of every gene sequence, the method reported by Yang and Nielsen [109] was adopted. Adaptive evolution of genes was confirmed by computing likelihood ratio tests (LRTs). The KaKs-calculator (Chinese Academy of Sciences, Beijing, China) [110] was also used to calculate K_A, K_S, and the K_A/K_S ratio, based on a model-averaging method.

4.7. Codon Usage Bias and Unconventional Initiation Codon

Codon usage and RSCU values [111] were estimated for all exons in the consensus protein-coding genes with the CodonW v1.4.2 program (University of Nottingham, Nottingham, UK) [112]. For the purpose of verifying the existence of unconventional initiation codon, we designed pairs of primers (Table S1) for polymerase chain reaction (PCR) amplification of target fragments within four species (three new species and *G. pentaphyllum*). The products of PCR were tested using 1% agarose gel electrophoresis and sequenced. The obtained sequences were mapped to the corresponding CPGs using the software GENEIOUS R8 (Biomatters Ltd., Auckland, New Zealand). Due to the lack of RNA sequence data, we blasted a 101 bp CPG sequence fragment of *G. pentaphyllum*, including the start codon of *psbL* gene, in the published transcriptome dataset of *G. pentaphyllum* (accession number in NCBI: SRX1364750) [113] using the online program BLASTn (https://blast.ncbi.nlm.nih.gov; U.S. National Library of Medicine, Bethesda, Rockville, MD, USA).

4.8. Phylogenetic Relationships

To reconstruct the phylogenetic relationship, 15 published complete CPG sequences from Cucurbitales were also selected in the analyses (Table S2). In total, 27 sequences were aligned using the MAFFT v7.017 program (Osaka University, Suita, Japan) [101]. Due to the differentiation of the molecular evolutionary rate among the different CPG regions, phylogenetic relationship analyses were performed using the following five datasets: (1) the overall CPG sequences; (2) the large-single copy region (LSC); (3) the small single-copy region (SSC); (4) one inverted repeats region (IRb); (5) consensus protein coding genes (CDS); and (6) eight highly variable regions (HVR). The best-fitting model for each dataset was determined by software Modeltest 3.7 (Brigham Young University, Provo, UT, USA) [114] under the Akaike information criterion. Bayesian inference (BI) was performed by MrBayes 3.12 (SwedishMuseum of Natural History, Stockholm, Sweden) [115] using the following parameters: Markov chain Monte Carlo simulations algorithm (MCMC) for 1×10^5 generations with four incrementally-heated chains. The maximum likelihood (ML) trees were implemented with RAxML v7.2.8 (Heidelberg Institute for Theoretical Studies, Heidelberg, Germany) [116] with 1000 replicates. In all analyses, *C. laevigata* and *N. tabacum* (Z00044) were chosen as outgroups.

5. Conclusions

The comparative analyses of complete CPGs contribute towards understanding the complete CPG structure and evolution, the identification of species, and the determination of phylogenetic relationships. Here, we have successfully applied Illumina sequencing to determine the complete CPGs of three herbaceous plants from the Cucurbitaceae, further enriching the valuable resources for the complete CPGs of higher plants. The results revealed that they shared most of the common genomic features with other species of Cucurbitaceae. Sequence divergence analysis showed high conservatism of the coding and IR regions. The coding proteins of three selective genes (*accD*, *clpP* and *matK*) were screened out, and they would contribute to analyzing the adaptive evolution. Evidence for RNA editing was demonstrated involving an unconventional initiation codon in the *psbL* gene. Phylogenetic analyses revealed that the genera *Gomphogyne*, *Hemsleya*, and *Gynostemma* were the earliest diverging lineage in Cucurbitaceae. The study suggested that the complete chloroplast genome sequences were useful for phylogenetic studies. This would enrich the valuable complete chloroplast genome resources of Cucurbitaceae, and determine potential SSR molecular markers and candidate DNA barcodes for coming phylogenetic and evolutionary population studies.

Supplementary Materials: The following are available online. Figure S1. Genome rearrangement events of 10 Cucurbitaceae species, comparing with *C. laevigata* and *N. tabacum*; Figure S2. Sequence identity plots among 11 chloroplast genomes, with *G. cissiformis* var. *cissiformis* as a reference by using mVISTA. The vertical scale indicates the identity percentage ranging from 50% to 100%. The horizontal axis corresponds to the coordinates within the chloroplast genome. Coding and non-coding regions are marked in blue and pink, respectively. Annotated genes are displayed along the top. The black boxes show the two IR regions; Figure S3. Amino acid frequencies and base compositions of chloroplast genomes of four species. (A) Frequency of amino acids usage; (B) Base compositions of chloroplast genomes; Figure S4. Phylogenetic relationship of the 27 species inferred from ML and BI analyses based on four datasets (A. LSC; B. SSC; C. IR; D. CDS; E. 8 HVR). The bootstrap values of ML analyses and Bayesian posterior probabilities are shown beside the clades. *C. laevigata* and *N. tabacum* were used as the outgroups; Table S1. Primers for low coverage regions of Cucurbitaceae species; Table S2. List of species and their basic information included in the analyses of complete chloroplast genomes; Table S3. Percentages of variable characters in coding and non-coding regions and the percentage of indels in variable characters; Table S4. PCR primers for relatively highly variable regions; Table S5. List of tandem repeats in the chloroplast genome of eleven Cucurbitales species; Table S6. List of dispersed repeats and palindromic repeats in the chloroplast genome of eleven Cucurbitales species; Table S7. List of SSRs in the chloroplast genome of eleven Cucurbitales species; Table S8. Summary of three categories of repeats and simple sequence repeats (SSRs); Table S9. Results of selective events analysis in KaKs-calculator; Table S10. Codon usage and relative synonymous codon usage (RSCU) value for protein-coding genes in the ten Cucurbitaceae chloroplast genomes; Table S11. Base compositions for protein-coding genes in the ten Cucurbitaceae chloroplast genomes.

Author Contributions: Conceptualization: X.Z. and G.Z.; Formal analysis: X.Z. and T.Z.; Funding acquisition: G.Z.; Investigation: X.Z., T.Z., J.Y., J.S., M.J. and Y.Z.; Methodology: T.Z.; Project administration: G.Z.; Resources: J.Y. and J.S.; Software: X.Z. and T.Z.; Supervision: J.Y. and G.Z.; Validation: X.Z.; Visualization: X.Z.; Writing—original draft: X.Z.; Writing—review and editing: T.Z., J.Y., J.S., M.J., Y.Z. and G.Z.

Funding: This work was supported by the National Natural Science Foundation of China (grant numbers 31270364 and 31770229) and the Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT) (grant number IRT1174).

Acknowledgments: We are grateful to Mo Yang from Katholieke Universiteit Leuven for the support and help in language improvement of the manuscript.

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

References

1. Schaefer, H.; Heibl, C.; Renner, S.S. Gourds afloat: A dated phylogeny reveals an asian origin of the gourd family (cucurbitaceae) and numerous oversea dispersal events. *Proc. Roy. Soc. B-Biol. Sci.* **2009**, *276*, 843–851. [CrossRef] [PubMed]
2. Lu, A.; Huang, L.; Chen, S.; Charles, J. *Flora of China*; Missouri Botanical Garden Press: Beijing, China, 2011; Volume 19.
3. 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] [PubMed]
4. Heiser, C.B. *The Gourd Book*; University of Oklahoma Press: Norman, OK, USA, 2016.
5. Tsai, Y.C.; Lin, C.L.; Chen, B.H. Preparative chromatography of flavonoids and saponins in gynostemma pentaphyllum and their antiproliferation effect on hepatoma cell. *Phytomedicine* **2010**, *18*, 2–10. [CrossRef] [PubMed]
6. Xie, Z.; Liu, W.; Huang, H.; Slavin, M.; Zhao, Y.; Whent, M.; Blackford, J.; Lutterodt, H.; Zhou, H.; Chen, P. Chemical composition of five commercial gynostemma pentaphyllum samples and their radical scavenging, antiproliferative, and anti-inflammatory properties. *J. Agric. Food Chem.* **2010**, *58*, 11243–11249. [CrossRef] [PubMed]
7. Nie, R.; Chen, Z. The research history and present status on the chemical components of genus hemsleya (cucurbitaceae). *Acta Bot. Yunnanica* **1986**, *8*, 115–124.
8. Dinan, L.; Whiting, P.; Sarker, S.D.; Kasai, R.; Yamasaki, K. Cucurbitane-type compounds from hemsleya carnosiflora antagonize ecdysteroid action in the drosophila melanogaster bii cell line. *Cell Mol. Life Sci.* **1997**, *53*, 271–274. [CrossRef] [PubMed]
9. Chiu, M.; Gao, J. Three new cucurbitacins from hemsleya lijiangensis. *Chin. Chem. Lett.* **2003**, *14*, 389–392.
10. Wu, J.; Wu, Y.; Yang, B.B. Anticancer activity of hemsleya amabilis extract. *Life Sci.* **2002**, *71*, 2161–2170. [CrossRef]

11. Jeffrey, C. A new system of cucurbitaceae. *Bot. Zhum.* **2005**, *90*, 332–335.
12. Piperno, D.R.; Stothert, K.E. Phytolith Evidence for Early Holocene *Cucurbita* Domestication in Southwest Ecuador. *Science* **2003**, *299*, 1054–1057. [CrossRef] [PubMed]
13. Kistler, L.; Montenegro, Á.; Smith, B.D.; Gifford, J.A.; Green, R.E.; Newsom, L.A.; Shapiro, B. Transoceanic drift and the domestication of african bottle gourds in the americas. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 2937. [CrossRef] [PubMed]
14. Guillaume, C.; Renner, S.S. Watermelon origin solved with molecular phylogenetics including linnaean material: Another example of museomics. *New Phytol.* **2015**, *205*, 526–532.
15. Kistler, L.; Newsom, L.A.; Ryan, T.M.; Clarke, A.C.; Smith, B.D.; Perry, G.H. Gourds and squashes (*Cucurbita* spp.) adapted to megafaunal extinction and ecological anachronism through domestication. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 15107–15112. [CrossRef] [PubMed]
16. Sanjur, O.I.; Piperno, D.R.; Andres, T.C.; Wessel-Beaver, L. Phylogenetic relationships among domesticated and wild species of *Cucurbita* (Cucurbitaceae) inferred from a mitochondrial gene: Implications for crop plant evolution and areas of origin. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 535–540. [CrossRef] [PubMed]
17. Clarke, A.C.; Burtenshaw, M.K.; McLenachan, P.A.; Erickson, D.L.; Penny, D. Reconstructing the origins and dispersal of the polynesian bottle gourd (lagenaria siceraria). *Mol. Biol. Evol.* **2006**, *23*, 893–900. [CrossRef] [PubMed]
18. Schaefer, H.; Renner, S.S. Phylogenetic relationships in the order cucurbitales and a new classification of the gourd family (cucurbitaceae). *Taxon* **2011**, *60*, 122–138.
19. Li, H.; Li, D. Systematic position of gomphogyne (cucurbitaceae) inferred from its, rpl16 and trns-trnr DNA sequences. *J. Syst. Evol.* **2008**, *46*, 595–599.
20. Li, H.-T.; Yang, J.-B.; Li, D.-Z.; Möller, M.; Shah, A. A molecular phylogenetic study of hemsleya (cucurbitaceae) based on its, rpl16, trnh-psba, and trnl DNA sequences. *Plant Syst. Evol.* **2010**, *285*, 23–32. [CrossRef]
21. Barrett, C.F.; Baker, W.J.; Comer, J.R.; Conran, J.G.; Lahmeyer, S.C.; Leebens-Mack, J.H.; Li, J.; Lim, G.S.; Mayfield-Jones, D.R.; Perez, L.; et al. Plastid genomes reveal support for deep phylogenetic relationships and extensive rate variation among palms and other commelinid monocots. *New Phytol.* **2016**, *209*, 855–870. [CrossRef] [PubMed]
22. Ren, T.; Yang, Y.; Zhou, T.; Liu, Z.-L. Comparative plastid genomes of primula species: Sequence divergence and phylogenetic relationships. *Int. J. Mol. Sci.* **2018**, *19*, 1050. [CrossRef] [PubMed]
23. Edger, P.P.; Hall, J.C.; Harkess, A.; Tang, M.; Coombs, J.; Mohammadin, S.; Schranz, M.E.; Xiong, Z.; Leebens-Mack, J.; Meyers, B.C.; et al. Brassicales phylogeny inferred from 72 plastid genes: A reanalysis of the phylogenetic localization of two paleopolyploid events and origin of novel chemical defenses. *Am. J. Bot.* **2018**, *105*, 463–469. [CrossRef] [PubMed]
24. Palmer, J.D. Comparative organization of chloroplast genomes. *Annu. Rev. Genet.* **1985**, *19*, 325–354. [CrossRef] [PubMed]
25. Wicke, S.; Schneeweiss, G.M.; de Pamphilis, C.W.; Müller, K.F.; Quandt, D. The evolution of the plastid chromosome in land plants: Gene content, gene order, gene function. *Plant Mol. Biol.* **2011**, *76*, 273–297. [CrossRef] [PubMed]
26. Zhang, X.; Zhou, T.; Kanwal, N.; Zhao, Y.; Bai, G.; Zhao, G. Completion of eight gynostemma bl. (cucurbitaceae) chloroplast genomes: Characterization, comparative analysis, and phylogenetic relationships. *Front. Plant Sci.* **2017**, *8*, 1583. [CrossRef] [PubMed]
27. Wang, R.-J.; Cheng, C.-L.; Chang, C.-C.; Wu, C.-L.; Su, T.-M.; Chaw, S.-M. Dynamics and evolution of the inverted repeat-large single copy junctions in the chloroplast genomes of monocots. *BMC Evol. Biol.* **2008**, *8*, 36. [CrossRef] [PubMed]
28. Yang, Y.; Zhou, T.; Duan, D.; Yang, J.; Feng, L.; Zhao, G. Comparative analysis of the complete chloroplast genomes of five quercus species. *Front. Plant Sci.* **2016**, *7*, 959. [CrossRef] [PubMed]
29. Zhong, B.; Yonezawa, T.; Zhong, Y.; Hasegawa, M. Episodic evolution and adaptation of chloroplast genomes in ancestral grasses. *PLoS ONE* **2009**, *4*, e5297. [CrossRef] [PubMed]
30. Rabah, S.O.; Shrestha, B.; Hajrah, N.H.; Sabir, M.J.; Alharby, H.F.; Sabir, M.J.; Alhebshi, A.M.; Sabir, J.S.M.; Gilbert, L.E.; Ruhlman, T.A.; et al. *Passiflora* plastome sequencing reveals widespread genomic rearrangements. *J. Syst. Evol.* 2018. [CrossRef]

31. Diekmann, K.; Hodkinson, T.R.; Wolfe, K.H.; van den Bekerom, R.; Dix, P.J.; Barth, S. Complete chloroplast genome sequence of a major allogamous forage species, perennial ryegrass (*lolium perenne* L.). *DNA Res.* **2009**, *16*, 165–176. [CrossRef] [PubMed]
32. Doyle, J.J.; Davis, J.I.; Soreng, R.J.; Garvin, D.; Anderson, M.J. Chloroplast DNA inversions and the origin of the grass family (poaceae). *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 7722–7726. [CrossRef] [PubMed]
33. Timme, R.E.; Kuehl, J.V.; Boore, J.L.; Jansen, R.K. A comparative analysis of the Lactuca and Helianthus (Asteraceae) plastid genomes: Identification of divergent regions and categorization of shared repeats. *Am. J. Bot.* **2007**, *94*, 302–312. [CrossRef] [PubMed]
34. Weng, M.L.; Blazier, J.C.; Govindu, M.; Jansen, R.K. Reconstruction of the ancestral plastid genome in Geraniaceae reveals a correlation between genome rearrangements, repeats and nucleotide substitution rates. *Mol. Biol. Evol.* **2013**, *31*, 645–659. [CrossRef] [PubMed]
35. Vanin, E.F. Processed pseudogenes: Characteristics and evolution. *Annu. Rev. Genet.* **1985**, *19*, 253–272. [CrossRef] [PubMed]
36. Balakirev, E.S.; Ayala, F.J. Pseudogenes: Are they "junk" or functional DNA? *Annu. Rev. Genet.* **2003**, *37*, 123–151. [CrossRef] [PubMed]
37. Asif, H.; Khan, A.; Iqbal, A.; Khan, I.A.; Heinze, B.; Azim, M.K. The chloroplast genome sequence of *syzygium cumini* (L.) and its relationship with other angiosperms. *Tree Genet. Genomes* **2013**, *9*, 867–877. [CrossRef]
38. Nashima, K.; Terakami, S.; Nishitani, C.; Kunihisa, M.; Shoda, M.; Takeuchi, M.; Urasaki, N.; Tarora, K.; Yamamoto, T.; Katayama, H. Complete chloroplast genome sequence of pineapple (ananas comosus). *Tree Genet. Genomes* **2015**, *11*, 60. [CrossRef]
39. Do, H.D.K.; Kim, J.S.; Kim, J.-H. Comparative genomics of four liliales families inferred from the complete chloroplast genome sequence of veratrum patulum o. Loes. (melanthiaceae). *Gene* **2013**, *530*, 229–235. [CrossRef] [PubMed]
40. Sato, S.; Nakamura, Y.; Kaneko, T.; Asamizu, E.; Tabata, S. Complete structure of the chloroplast genome of arabidopsis thaliana. *DNA Res.* **1999**, *6*, 283–290. [CrossRef] [PubMed]
41. Maréchal, A.; Brisson, N. Recombination and the maintenance of plant organelle genome stability. *New Phytol.* **2010**, *186*, 299–317. [CrossRef] [PubMed]
42. Kim, K.-J.; Lee, H.-L. Complete chloroplast genome sequences from korean ginseng (panax schinseng nees) and comparative analysis of sequence evolution among 17 vascular plants. *DNA Res.* **2004**, *11*, 247–261. [CrossRef] [PubMed]
43. Hansen, D.R.; Dastidar, S.G.; Cai, Z.; Penaflor, C.; Kuehl, J.V.; Boore, J.L.; Jansen, R.K. Phylogenetic and evolutionary implications of complete chloroplast genome sequences of four early-diverging angiosperms: Buxus (buxaceae), chloranthus (chloranthaceae), dioscorea (dioscoreaceae), and illicium (schisandraceae). *Mol. Phylogenet. Evol.* **2007**, *45*, 547–563. [CrossRef] [PubMed]
44. Naydenov, K.D.; Naydenov, M.K.; Alexandrov, A.; Vasilevski, K.; Gyuleva, V.; Matevski, V.; Nikolic, B.; Goudiaby, V.; Bogunic, F.; Paitaridou, D. Ancient split of major genetic lineages of european black pine: Evidence from chloroplast DNA. *Tree Genet. Genomes* **2016**, *12*, 68. [CrossRef]
45. Nei, M.; Kumar, S. *Molecular Evolution and Phylogenetics*; Oxford University Press: Oxford, UK, 2000.
46. Yin, K.; Zhang, Y.; Li, Y.; Du, F. Different natural selection pressures on the atpf gene in evergreen sclerophyllous and deciduous oak species: Evidence from comparative analysis of the complete chloroplast genome of quercus aquifolioides with other oak species. *Int. J. Mol. Sci.* **2018**, *19*, 1042. [CrossRef] [PubMed]
47. Tomoko, O. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *J. Mol. Evol.* **1995**, *40*, 56–63. [CrossRef]
48. Kimura, M. The neutral theory of molecular evolution and the world view of the neutralists. *Genome* **1989**, *31*, 24–31. [CrossRef] [PubMed]
49. Sasaki, Y.; Nagano, Y. Plant acetyl-coa carboxylase: Structure, biosynthesis, regulation, and gene manipulation for plant breeding. *Biosci. Biotech. Bioch.* **2004**, *68*, 1175–1184. [CrossRef] [PubMed]
50. 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]
51. Wakasugi, T.; Tsudzuki, T.; Sugiura, M. The genomics of land plant chloroplasts: Gene content and alteration of genomic information by rna editing. *Photosynth. Res.* **2001**, *70*, 107–118. [CrossRef] [PubMed]

52. Tseng, C.-C.; Sung, T.-Y.; Li, Y.-C.; Hsu, S.-J.; Lin, C.-L.; Hsieh, M.-H. Editing of accd and ndhf chloroplast transcripts is partially affected in the arabidopsis vanilla cream1 mutant. *Plant Mol. Biol.* **2010**, *73*, 309–323. [CrossRef] [PubMed]
53. Madoka, Y.; Tomizawa, K.-I.; Mizoi, J.; Nishida, I.; Nagano, Y.; Sasaki, Y. Chloroplast transformation with modified accd operon increases acetyl-coa carboxylase and causes extension of leaf longevity and increase in seed yield in tobacco. *Plant Cell Physiol.* **2002**, *43*, 1518–1525. [CrossRef] [PubMed]
54. Katayama, H.; Ogihara, Y. Phylogenetic affinities of the grasses to other monocots as revealed by molecular analysis of chloroplast DNA. *Curr. Genet.* **1996**, *29*, 572–581. [CrossRef] [PubMed]
55. Shikanai, T.; Shimizu, K.; Ueda, K.; Nishimura, Y.; Kuroiwa, T.; Hashimoto, T. The chloroplast clpp gene, encoding a proteolytic subunit of atp-dependent protease, is indispensable for chloroplast development in tobacco. *Plant Cell Physiol.* **2001**, *42*, 264–273. [CrossRef] [PubMed]
56. Clarke, A.K.; Gustafsson, P.; Lidholm, J.Å. Identification and expression of the chloroplast clpp gene in the conifer pinus contorta. *Plant Mol. Biol.* **1994**, *26*, 851–862. [CrossRef] [PubMed]
57. Kuroda, H.; Maliga, P. The plastid clpp1 protease gene is essential for plant development. *Nature* **2003**, *425*, 86. [CrossRef] [PubMed]
58. Maliga, P. Plastid transformation in higher plants. *Annu. Rev. Plant Biol.* **2004**, *55*, 289–313. [CrossRef] [PubMed]
59. Zoschke, R.; Nakamura, M.; Liere, K.; Sugiura, M.; Börner, T.; Schmitz-Linneweber, C. An organellar maturase associates with multiple group ii introns. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 3245–3250. [CrossRef] [PubMed]
60. Neuhaus, H.; Link, G. The chloroplast trnalys(uuu) gene from mustard (sinapis alba) contains a class ii intron potentially coding for a maturase-related polypeptide. *Curr. Genet.* **1987**, *11*, 251–257. [CrossRef] [PubMed]
61. Mohr, G.; Perlman, P.S.; Lambowitz, A.M. Evolutionary relationships among group ii intron-encoded proteins and identification of a conserved domain that may be related to maturase function. *Nucleic Acids Res.* **1993**, *21*, 4991–4997. [CrossRef] [PubMed]
62. Johnson, L.A.; Soltis, D.E. Phylogenetic inference in saxifragaceae sensu stricto and gilia (polemoniaceae) using matk sequences. *Ann. Mo. Bot. Gard.* **1995**, *82*, 149–175. [CrossRef]
63. Gadek, P.A.; Wilson, P.G.; Quinn, C.J. Phylogenetic reconstruction in myrtaceae using mat k, with particular reference to the position of psiloxylon and heteropyxis. *Aust. Syst. Bot.* **1996**, *9*, 283–290. [CrossRef]
64. Hilu, K.; Liang, H. The matk gene: Sequence variation and application in plant systematics. *Am. J. Bot.* **1997**, *84*, 830–839. [CrossRef] [PubMed]
65. Sueoka, N.; Kawanishi, Y. DNA G + C content of the third codon position and codon usage biases of human genes. *Gene* **2000**, *261*, 53–62. [CrossRef]
66. Bellgard, M.; Schibeci, D.; Trifonov, E.; Gojobori, T. Early detection of G + C differences in bacterial species inferred from the comparative analysis of the two completely sequenced helicobacter pylori strains. *J. Mol. Evol.* **2001**, *53*, 465–468. [CrossRef] [PubMed]
67. Shimda, H.; Sugiuro, M. Fine structural features of the chloroplast genome: Comparison of the sequenced chloroplast genomes. *Nucleic Acids Res.* **1991**, *19*, 983–995. [CrossRef]
68. Clegg, M.T.; Gaut, B.S.; Learn, G.H., Jr.; Morton, B.R. Rates and patterns of chloroplast DNA evolution. *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 6795–6801. [CrossRef] [PubMed]
69. Tangphatsornruang, S.; Sangsrakru, D.; Chanprasert, J.; Uthaipaisanwong, P.; Yoocha, T.; Jomchai, N.; Tragoonrung, S. The chloroplast genome sequence of mungbean (vigna radiata) determined by high-throughput pyrosequencing: Structural organization and phylogenetic relationships. *DNA Res.* **2009**, *17*, 11–22. [CrossRef] [PubMed]
70. Delannoy, E.; Fujii, S.; Colas, d.F.-S.C.; Brundrett, M.; Small, I. Rampant gene loss in the underground orchid rhizanthella gardneri highlights evolutionary constraints on plastid genomes. *Mol. Biol. Evol.* **2011**, *28*, 2077–2086. [CrossRef] [PubMed]
71. Yang, Y.; Zhu, J.; Feng, L.; Zhou, T.; Bai, G.; Yang, J.; Zhao, G. Plastid genome comparative and phylogenetic analyses of the key genera in fagaceae: Highlighting the effect of codon composition bias in phylogenetic inference. *Front. Plant Sci.* **2018**, *9*, 82. [CrossRef] [PubMed]
72. Tsudzuki, T.; Wakasugi, T.; Sugiura, M. Comparative analysis of rna editing sites in higher plant chloroplasts. *J. Mol. Evol.* **2001**, *53*, 327–332. [CrossRef] [PubMed]

73. Maier, R.M.; Zeltz, P.; Kössel, H.; Bonnard, G.; Gualberto, J.M.; Grienenberger, J.M. Rna editing in plant mitochondria and chloroplasts. *Plant Mol. Biol.* **1996**, *32*, 343–365. [CrossRef] [PubMed]
74. Hoch, B.; Maier, R.M.; Appel, K.; Igloi, G.L.; Kössel, H. Editing of a chloroplast mrna by creation of an initiation codon. *Nature* **1991**, *353*, 178–180. [CrossRef] [PubMed]
75. Freyer, R.; Kiefer-Meyer, M.C.; Kössel, H. Occurrence of plastid rna editing in all major lineages of land plants. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 6285–6290. [CrossRef] [PubMed]
76. Tillich, M.; Funk, H.L.C.; Poltnigg, P.; Sabater, B.; Martin, M.; Maier, R. Editing of plastid rna in arabidopsis thaliana ecotypes. *Plant J.* **2010**, *43*, 708–715. [CrossRef] [PubMed]
77. Maier, R.M.; Neckermann, K.; Igloi, G.L.; Kössel, H. Complete sequence of the maize chloroplast genome: Gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. *J. Mol. Biol.* **1995**, *251*, 614–628. [CrossRef] [PubMed]
78. Corneille, S.; Lutz, K.; Maliga, P. Conservation of rna editing between rice and maize plastids: Are most editing events dispensable? *Mol. Gen. Genet.* **2000**, *264*, 419–424. [CrossRef] [PubMed]
79. Guzowska-Nowowiejska, M.; Fiedorowicz, E.; Pląder, W. Cucumber, melon, pumpkin, and squash: Are rules of editing in flowering plants chloroplast genes so well known indeed? *Gene* **2009**, *434*, 1–8. [CrossRef] [PubMed]
80. Wakasugi, T.; Hirose, T.; Horihata, M.; Tsudzuki, T.; Kössel, H.; Sugiura, M. Creation of a novel protein-coding region at the RNA level in black pine chloroplasts: The pattern of RNA editing in the gymnosperm chloroplast is different from that in angiosperms. *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 8766–8770. [CrossRef] [PubMed]
81. Yoshinaga, K.; Kakehi, T.; Shima, Y.; Iinuma, H.; Masuzawa, T.; Ueno, M. Extensive rna editing and possible double-stranded structures determining editing sites in the atpb transcripts of hornwort chloroplasts. *Nucleic Acids Res.* **1997**, *25*, 4830–4834. [CrossRef] [PubMed]
82. Ozawa, S.; Kobayashi, T.; Sugiyama, R.; Hoshida, H.; Shiina, T.; Toyoshima, Y. Role of psii-l protein (*psbl* gene product) on the electron transfer in photosystem ii complex. 1. Over-production of wild-type and mutant versions of psii-l protein and reconstitution into the psii core complex. *Plant Mol. Biol.* **1997**, *34*, 151–161. [CrossRef] [PubMed]
83. Kuntz, M.; Camara, B.; Weil, J.H.; Schantz, R. The *psbl* gene from bell pepper (capsicum annuum): Plastid rna editing also occurs in non-photosynthetic chromoplasts. *Plant Mol. Biol.* **1992**, *20*, 1185–1188. [CrossRef] [PubMed]
84. Kudla, J.; Igloi, G.L.; Metzlaff, M.; Hagemann, R.; Kössel, H. Rna editing in tobacco chloroplasts leads to the formation of a translatable *psbl* mrna by a c to u substitution within the initiation codon. *Embo J.* **1992**, *11*, 1099–1103. [PubMed]
85. Bock, R.; Hagemann, R.; Kössel, H.; Kudla, J. Tissue- and stage-specific modulation of rna editing of the psbf and *psbl* transcript from spinach plastids—A new regulatory mechanism? *Mol. Gen. Genet.* **1993**, *240*, 238–244. [CrossRef] [PubMed]
86. Raman, G.; Park, S. The complete chloroplast genome sequence of ampelopsis: Gene organization, comparative analysis, and phylogenetic relationships to other angiosperms. *Front. Plant Sci.* **2016**, *7*, 341. [CrossRef] [PubMed]
87. Hirose, T.; Sugiura, M. Both RNA editing and RNA cleavage are required for translation of tobacco chloroplast *ndhDmRNA*: A possible regulatory mechanism for the expression of a chloroplast operon consisting of functionally unrelated genes. *Embo J.* **1997**, *16*, 6804–6811. [CrossRef] [PubMed]
88. Bock, D.G.; Kane, N.C.; Ebert, D.P.; Rieseberg, L.H. Genome skimming reveals the origin of the jerusalem artichoke tuber crop species: Neither from jerusalem nor an artichoke. *New Phytol.* **2014**, *201*, 1021–1030. [CrossRef] [PubMed]
89. Carbonell-Caballero, J.; Alonso, R.; Ibañez, V.; Terol, J.; Talon, M.; Dopazo, J. A phylogenetic analysis of 34 chloroplast genomes elucidates the relationships between wild and domestic species within the genuscitrus. *Mol. Biol. Evol.* **2015**, *32*, 2015–2035. [CrossRef] [PubMed]
90. Yu, X.Q.; Gao, L.M.; Soltis, D.E.; Soltis, P.S.; Yang, J.B.; Fang, L.; Yang, S.X.; Li, D.Z. Insights into the historical assembly of east asian subtropical evergreen broadleaved forests revealed by the temporal history of the tea family. *New Phytol.* **2017**, *215*, 1235–1248. [CrossRef] [PubMed]
91. Li, D.Z. *Phylogeny and Evolution of the Genus Hemsleya*; Yunnan Science and Technology Press: Kunming, China, 1993.

92. Doyle, J.J. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* **1987**, *19*, 11–15.
93. Zhu, Q.; Cui, H.; Zhao, Y.; Gao, P.; Liu, S.; Wang, P.; Luan, F. The complete chloroplast genome sequence of the *Citrullus lanatus* L. Subsp. Vulgaris (cucurbitaceae). *Mitochondrial DNA B* **2016**, *1*, 943–944. [CrossRef]
94. Sousa, A.; Bellot, S.; Fuchs, J.; Houben, A.; Renner, S.S. Analysis of transposable elements and organellar DNA in male and female genomes of a species with a huge y chromosome reveals distinct y centromeres. *Plant J.* **2016**, *88*, 387–396. [CrossRef] [PubMed]
95. Pląder, W.; Yukawa, Y.; Sugiura, M.; Malepszy, S. The complete structure of the cucumber (*Cucumis sativus* L.) chloroplast genome: Its composition and comparative analysis. *Cell. Mol. Biol. Lett.* **2007**, *12*, 584–594. [CrossRef] [PubMed]
96. Hahn, C.; Bachmann, L.; Chevreux, B. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—A baiting and iterative mapping approach. *Nucleic Acids Res.* **2013**, *41*, e129. [CrossRef] [PubMed]
97. Chevreux, B.; Pfisterer, T.; Drescher, B.; Driesel, A.J.; Müller, W.E.G.; Wetter, T.; Suhai, S. Using the miraEST assembler for reliable and automated mrna transcript assembly and SNP detection in sequenced ESTs. *Genome Res.* **2004**, *14*, 1147–1159. [CrossRef] [PubMed]
98. Untergasser, A.; Cutcutache, I.; Koressaar, T.; Ye, J.; Faircloth, B.C.; Remm, M.; Rozen, S.G. Primer 3—New capabilities and interfaces. *Nucleic Acids Res.* **2012**, *40*, e115. [CrossRef] [PubMed]
99. Wyman, S.K.; Jansen, R.K.; Boore, J.L. Automatic annotation of organellar genomes with dogma. *Bioinformatics* **2004**, *20*, 3252–3255. [CrossRef] [PubMed]
100. Frazer, K.A.; Pachter, L.; Poliakov, A.; Rubin, E.M.; Dubchak, I. Vista: Computational tools for comparative genomics. *Nucleic Acids Res.* **2004**, *32*, W273–W279. [CrossRef] [PubMed]
101. Katoh, K.; Standley, D.M. Mafft multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [CrossRef] [PubMed]
102. Librado, P.; Rozas, J. Dnasp v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **2009**, *25*, 1451–1452. [CrossRef] [PubMed]
103. Tamura, K.; Peterson, D.; Peterson, N.; Stecher, G.; Nei, M.; Kumar, S. Mega5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **2011**, *28*, 2731–2739. [CrossRef] [PubMed]
104. Zhang, Y.-J.; Ma, P.-F.; Li, D.-Z. High-throughput sequencing of six bamboo chloroplast genomes: Phylogenetic implications for temperate woody bamboos (poaceae: Bambusoideae). *PLoS ONE* **2011**, *6*, e20596. [CrossRef] [PubMed]
105. Thiel, T.; Michalek, W.; Varshney, R.; Graner, A. Exploiting est databases for the development and characterization of gene-derived ssr-markers in barley (*Hordeum vulgare* L.). *Theor. Appl. Genet.* **2003**, *106*, 411–422. [CrossRef] [PubMed]
106. Benson, G. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* **1999**, *27*, 573–580. [CrossRef] [PubMed]
107. 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]
108. Yang, Z. Paml 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **2007**, *24*, 1586–1591. [CrossRef] [PubMed]
109. Yang, Z.; Nielsen, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **2000**, *17*, 32–43. [CrossRef] [PubMed]
110. Zhang, Z.; Li, J.; Zhao, X.-Q.; Wang, J.; Wong, G.K.-S.; Yu, J. Kaks_calculator: Calculating ka and ks through model selection and model averaging. *Genom. Proteom. Bioinf.* **2006**, *4*, 259–263. [CrossRef]
111. Sharp, P.M.; Li, W.-H. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **1987**, *15*, 1281–1295. [CrossRef] [PubMed]
112. Peden, J.F. Analysis of Codon Usage. Ph.D. Thesis, University of Nottingham, Nottingham, UK, 1999.
113. Zhao, Y.-M.; Zhou, T.; Li, Z.-H.; Zhao, G.-F. Characterization of global transcriptome using illumina paired-end sequencing and development of est-ssr markers in two species of *Gynostemma* (Cucurbitaceae). *Molecules* **2015**, *20*, 21214–21231. [CrossRef] [PubMed]

114. Posada, D.; Crandall, K.A. Modeltest: Testing the model of DNA substitution. *Bioinformatics* **1998**, *14*, 817–818. [CrossRef] [PubMed]

115. Ronquist, F.; Huelsenbeck, J.P. Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **2003**, *19*, 1572–1574. [CrossRef] [PubMed]

116. Stamatakis, A. Raxml-vi-hpc: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **2006**, *22*, 2688–2690. [CrossRef] [PubMed]

Sample Availability: Samples of the compounds are available from the authors.