



OPEN

Comparative chloroplast genome analysis of seven extant *Citrullus* species insight into genetic variation, phylogenetic relationships, and selective pressure

Cong Zhou^{1,4}, Putao Wang^{1,4}, Qun Zeng¹, Rongbin Zeng¹, Wei Hu¹, Lei Sun³, Shi Liu², Feishi Luan² & Qianglong Zhu¹✉

Citrullus ecirrhosus, *Citrullus rehmii*, and *Citrullus naudinianus* are three important related wild species of watermelon in the genus *Citrullus*, and their morphological differences are clear, however, their chloroplast genome differences remain unknown. This study is the first to assemble, analyze, and publish the complete chloroplast genomes of *C. ecirrhosus*, *C. rehmii*, and *C. naudinianus*. A comparative analysis was then conducted among the complete chloroplast genomes of seven extant *Citrullus* species, and the results demonstrated that the average genome sizes of *Citrullus* is 157,005 bp, a total of 130–133 annotated genes were identified, including 8 rRNA, 37 tRNA and 85–88 protein-encoding genes. Their gene content, order, and genome structure were similar. However, noncoding regions were more divergent than coding regions, and *rps16-trnQ* was a hypervariable fragment. Thirty-four polymorphic SSRs, 1,271 SNPs and 234 INDELS were identified. Phylogenetic trees revealed a clear phylogenetic relationship of *Citrullus* species, and the developed molecular markers (SNPs and *rps16-trnQ*) could be used for taxonomy in *Citrullus*. Three genes (*atpB*, *clpP1*, and *rpoC2*) were identified to undergo selection and would promote the environmental adaptation of *Citrullus*.

The genus *Citrullus* contains approximately seven extant species, *Citrullus lanatus*, *Citrullus mucospermus*, *Citrullus amarus*, *Citrullus ecirrhosus*, *Citrullus rehmii*, *Citrullus colocynthis*, and *Citrullus naudinianus*^{1,2}. *C. lanatus* subsp. *vulgaris* (Watermelon) is a worldwide cultivated horticultural crop bearing large fruits with crisp sweet, juicy, and red flesh and a green rind with black stripes. Its fruit often contains over 90% water and many nutritional compounds, including sugars, carotenoids and amino acids for cardiovascular prevention³, which makes it one of the most popular fruit crops around the world. Recent research suggested that the center of diversity and origin of modern watermelon, which is now cultivate and consume all over the world, is Kordofan in Sudan². In addition to *Citrullus lanatus* subsp. *vulgaris*, the genus *Citrullus* comprises six other cultivated or wild species, *C. naudinianus* is the most primitive wild species, and it is dieicious and widely distributed in sub-Saharan Africa⁴. *C. colocynthis* is used for folk medicine with therapeutic effects for various diseases, such as antidiabetic and anticancer effects^{5–7}, and mainly originates from northern Africa. *C. rehmii* and *C. ecirrhosus* commonly grow in desert climates and environments, and they are commonly found in southern Africa¹. *C. amarus* is a citron watermelon that can be directly used as a water source and be processed into jam and feed-stuff, and it originated from southern Africa and spread to the Mediterranean region⁸. *C. mucospermus* is an

¹Jiangxi Key Laboratory of Crop Physiology, Ecology and Genetic Breeding, Jiangxi Agricultural University, NO. 1101 Zhimin Street, Qingshanhu District, Nanchang 330045, People's Republic of China. ²College of Horticulture and Landscape Architecture, Northeast Agricultural University, Harbin 150030, People's Republic of China. ³Department of Agronomy and Horticulture, Liaoning Agricultural Technical College, Yingkou 115009, People's Republic of China. ⁴These authors contributed equally: Cong Zhou and Putao Wang. ✉email: longzhu2011@126.com

egusi watermelon with large edible seed and is native to western Africa². Furthermore, these wild species have been broadly applied to explore phylogenetic relationships and genetically improve cultivated watermelon (*C. lanatus* subsp. *vulgaris*)^{1,2,8,9}.

The chloroplast is an essential semi-autonomous organelle that mainly conducts photosynthesis and is important for the growth and development of plants and green algae. Chloroplasts are also responsible for the synthesis of starch, fatty acids, amino acids and other metabolites, and play an important role in the stress response^{10–12}. Chloroplasts possess a circular double-stranded DNA molecule. The complete chloroplast genome (CPG) usually demonstrates a relatively conserved quadripartite structure including a large single copy (LSC) and a small single copy (SSC) region, and two inverted repeats (IRa and IRb). Compared with nuclear genomes, the chloroplast genome has a unique pattern of maternal inheritance, a large number of genes and a lower mutation rate in adaptive evolution. Therefore, CPGs have been proposed as molecular markers to identify and classify plant species by the Consortium for the Barcode of Life (CBOL)¹³. The first chloroplast genome sequencing of tobacco was completed in 1986. By September 2022, there were 8,210 complete plant chloroplast or plastid genomes deposited in genome database of the National Center for Biotechnology Information (NCBI), comprising ~315 genes (encoding proteins, transfer RNA and ribosomal RNA) ranging from 15 to 707 kb in size. The chloroplast reference genomes of *C. lanatus* subsp. *vulgaris*, *C. amarus* and *C. colocynthis* have been published in our past researches^{14–16}. And the chloroplast reference genome of *C. mucospermus* was also published in 2017¹⁷. To explore the phylogenetic relationship within the genus *Citrullus* and especially the genetic variation of chloroplast genomes during the selection and domestication of cultivated watermelon, there is an urgent need for more chloroplast genomes of other *Citrullus* species to be published.

Here, we aimed to assemble and analyze the chloroplast genomes of three wild watermelon species, *C. rehmii*, *C. ecirrhosus* and *C. naudinianus*, based on their whole genome sequence data stored in the Sequence Read Archive (SRA) database of NCBI. We next compared their chloroplast genomes with the published CPGs of four other *Citrullus* species. Our study aims were: (1) to publish the CPGs of *C. ecirrhosus*, *C. rehmii*, and *C. naudinianus*, (2) to reveal the structural features of seven CPGs and evaluate the genetic variation in the chloroplast genome of *Citrullus*, and (3) to identify the molecular markers and the genes under positive selection, explore the phylogenetic relationships of seven *Citrullus* species, with the hope of providing insights into the speciation, selection and domestication of cultivated watermelon.

Results

Chloroplast genome generation, annotation, and comparative feature analysis. The sequence data (0.8 Gb) of *C. ecirrhosus*, *C. rehmii*, and *C. naudinianus* were extracted randomly, and approximately 13.27% (*C. ecirrhosus*), 21.37% (*C. rehmii*) and 12.75% (*C. naudinianus*) of these clean paired reads mapped to the CPG of *C. lanatus* subsp. *vulgaris*. The CPGs of these three *Citrullus* species were obtained after de novo assembly and strict validation. Then, they were annotated and archived in GenBank with accession number ON597627, MZ577999, and MZ577998.

To reveal the structural features of seven CPGs in the genus *Citrullus*, comparative analysis was conducted (Table 1). The average length of the CPGs is 157,005 bp, ranging from 156,906 bp (*C. lanatus* subsp. *vulgaris*) to 157,147 bp (*C. colocynthis*), and they demonstrated a typical quadripartite structure, including the LSC (86,728–87,007 bp) and SSC (17,897–17,999 bp) regions, and a pair of IRs (26,072–26,149 bp), but the variation coefficient of the SSC regions (2.57E–3) was greater than that of the LSC regions (9.56E–4) and IR regions

Genome features	<i>C. ecirrhosus</i> *	<i>C. naudinianus</i> *	<i>C. rehmii</i> *	<i>C. lanatus</i> subsp. <i>vulgaris</i>	<i>C. colocynthis</i>	<i>C. amarus</i>	<i>C. mucospermus</i>	Average ± SD	variation coefficient
Size (bp)	157,009	156,926	157,071	156,906	157,147	157,008	156,907	157,005 ± 84	5.36E–4
LSC length (bp)	86,823	86,728	86,887	86,845	86,852	86,813	86,846	86,845 ± 46	5.32E–4
SSC length (bp)	17,890	17,908	17,886	17,897	17,999	17,897	17,897	17,925 ± 36	2.05E–3
IR length (bp)	26,148	26,145	26,149	26,082	26,148	26,149	26,082	26,118 ± 29	1.13E–3
Genes number	133	133	133	131	131	130	131	132	0
Protein-coding genes	88	88	88	86	86	85	86	87	0
tRNA	37	37	37	37	37	37	37	37	0
rRNA	8	8	8	8	8	8	8	8	0
Overall GC content (%)	37.18	37.12	37.16	37.18	37.14	37.17	37.18	37.16 ± 0.02	–
GC content in LSC (%)	34.96	34.89	34.94	34.94	34.90	34.95	34.95	34.93 ± 0.03	–
GC content in SSC (%)	31.53	31.43	31.52	31.54	31.47	31.52	31.54	31.51 ± 0.04	–
GC content in IR (%)	42.80	42.76	42.80	42.83	42.82	42.81	42.83	42.81 ± 0.02	–
Accession no. in GenBank	MZ577999	MZ577998	ON597627	NC_032008/ KY014105	NC_035727/ MF357889	NC_035974/ MF536694	NC_033899/ KY430686	–	–

Table 1. Summary statistics for the assembly of seven CPGs in the genus *Citrullus*. LSC Large single copy, SSC small single copy, IR inverted repeat. *Indicated the chloroplast genome of species was assembled in the study.

($1.42E-3$). The average GC content (%) in the seven CPGs was approximately 37.16. Further analysis of GC content found that the IR region had the highest GC content compared to any other region in each CPG (Table 1). When the gene content of chloroplast genomes was compared, we found some mistakes in the original annotation of *C. lanatus* subsp. *vulgaris* and *C. mucospermus*, then we updated the chloroplast genome annotation of *C. lanatus* subsp. *vulgaris* at NCBI. After correcting the chloroplast genome annotation of *C. lanatus* subsp. *vulgaris* and *C. mucospermus*, we found that the seven CPGs encoded a set of 130–133 functional genes, including 85–88 protein-encoding genes, 37 transfer RNAs (tRNA), and 8 ribosomal RNAs (rRNA). Furthermore, 79 protein-encoding genes were identified among seven *Citrullus* species. Twenty-three genes were found to harbor introns, four of which (two *rps12* gene, one *clpP* gene, and one *ycf3* gene) had two introns. A consensus radius map of the seven CPGs was plotted to explore the variation in the IR/SC boundary (Supplementary Fig. S1), suggested that there was no the significant variation in the IR/SC boundary. Moreover, fine collinearity was observed among the seven CPGs of *Citrullus* species (Supplementary Fig. S2). Based on the conclusion that the seven CPGs have similar total gene numbers, names, and orders, their annotated CPGs of the seven species were drawn in a circle map (Fig. 1).

Chloroplast genome sequence divergence and nucleotide diversity analysis. The CPGs of *C. rehmii*, *C. ecirrhosus*, and *C. naudinianus* were aligned with those of four other four *Citrullus* species using

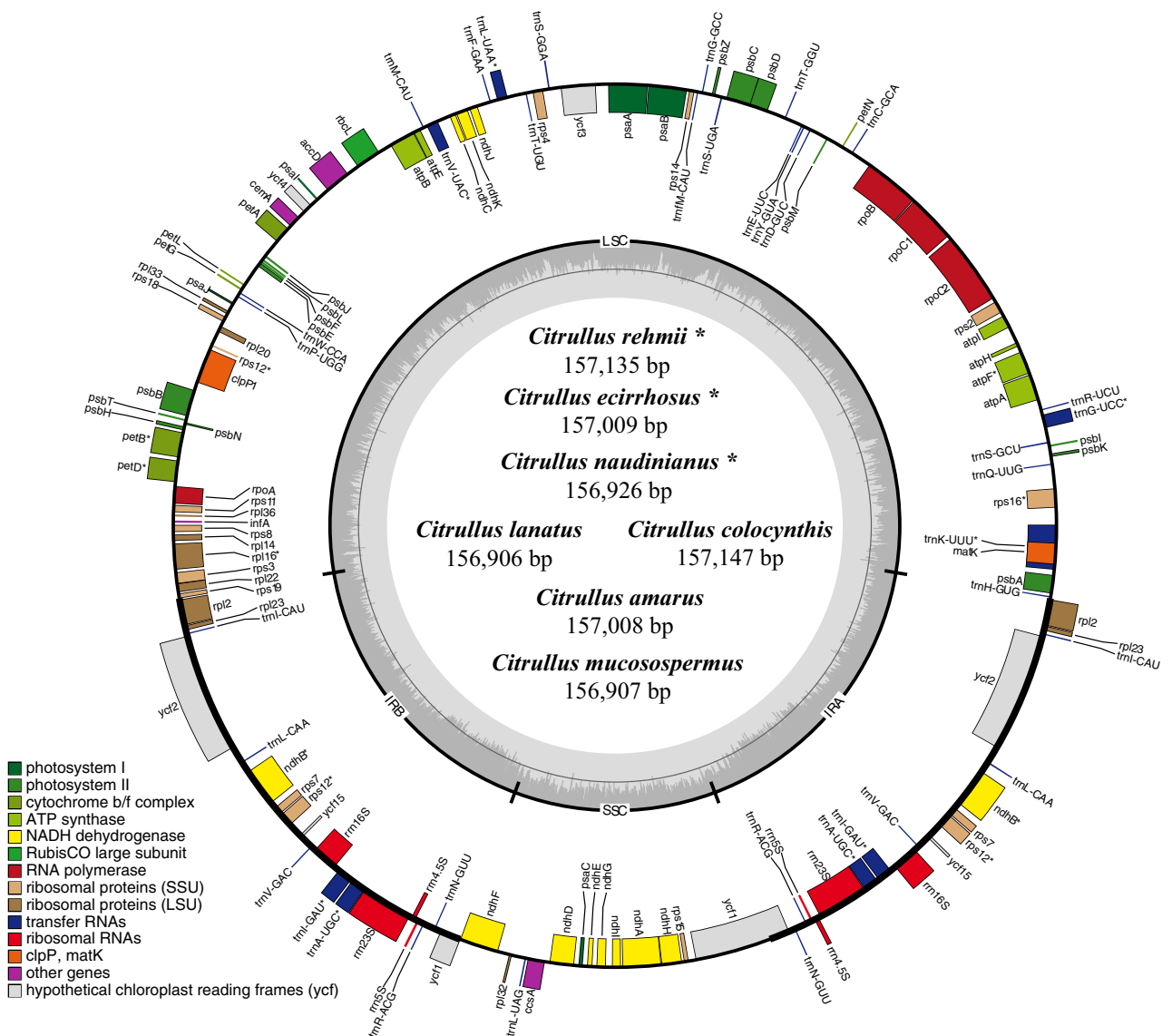


Figure 1. Complete chloroplast genome circle map of the seven studied *Citrullus* species. Genes outside the circle are transcribed on the positive chain, whereas those genes inside the circle are transcribed on the negative chain. Boxes with different colors represent genes with a different functional group. The inner ring shows the positions of LSC, IRa, SSC, and IRb, with the GC clamp of the chloroplast genome shown in grey. The chloroplast genome map was draw using OGDRAW v1.3.1 (<https://chlorobox.mpimp-golm.mpg.de/OGDraw.html>).

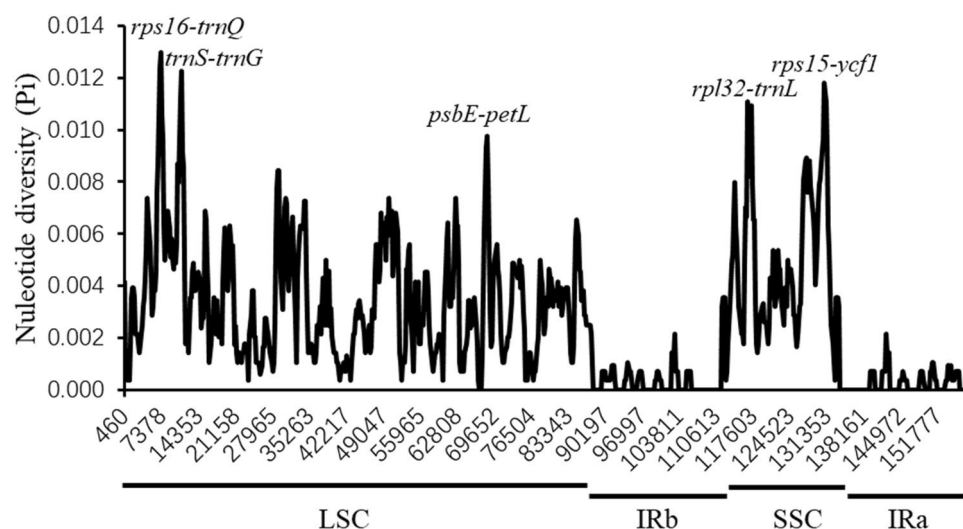


Figure 3. Position of the midpoint in the window Sliding window plots of nucleotide diversity (π) across the complete chloroplast genome of the seven *Citrullus* species. Window length: 800 bp; Step size: 200 bp.

DNA barcode	Location ^a	Aligned length	Pi	VS	PIn	K-2P distance	DSR (%)
<i>rps16-trnQ</i>	6538–7748	1222	0.013	40	9	0.011	100
<i>trnS-trnG</i>	10,397–11,437	1091	0.012	36	6	0.012	25
<i>psbE-petL</i>	67,227–68,042	816	0.010	24	4	0.010	50
<i>rpl32-trnL</i>	115,354–116,867	1556	0.011	40	11	0.010	75
<i>rps15-ycf1</i>	129,498–130,696	1199	0.012	40	6	0.011	75

Table 2. Comparative analysis of five hypervariable fragments in *Citrullus*. ^aThe location of five hypervariable fragments is confirmed based on the chloroplast genome of *C. lanatus* subsp. *vulgaris*. VS Variable sites, PIn Parsimony informative sites, DSR Discrimination success ratio.

Many repeat sequences were mainly distributed within the noncoding regions (intergenic and intron regions) of the seven CPGs (Supplementary Table S4).

The average SSR number was 56 (Supplementary Table S5). Mono-, di-, tri-, tetra-, penta-, and hexa-nucleotide SSRs in each species were analyzed, but only mono-, di-, tri-, and tetra-nucleotide SSRs were found (Fig. 4B), and the average number and proportions of mono-, di-, tri-, and tetra-nucleotide SSRs were 40(71.6%), 12(21.1%), 2(3.6%), and 2(3.6%), respectively (Supplementary Table S5). Notably, no penta- or hexa-nucleotide SSRs were found in the seven CPGs of *Citrullus*, and the number of mononucleotide SSRs in *C. naudinianus* (45) was the highest among the seven species. Nevertheless, the SSRs in these chloroplast genomes were remarkably rich in A/T content, with lower G/C content. The A/T mononucleotide repeats constituted a large portion of the SSRs in all species. In addition, we found that the seven species shared 47 SSRs, and 34 of those SSRs were polymorphic (Table S6). Three selected SSR markers were polymorphic among *C. lanatus* subsp. *vulgaris*, *C. mucospermus* and *C. amarus* (Fig. 5, Supplementary Fig. S4), suggesting that these SSR markers could be applied for species classification in *Citrullus*.

Chloroplast genome variation map. To identify the variation in the chloroplast genome between cultivated watermelon and the other six *Citrullus* species, 1271 single nucleotide polymorphisms (SNPs) were identified using VarScan, of which 769 (61.5%) SNPs were in noncoding regions, such as introns, and upstream and downstream regions of genes, and approximately 39.5% of SNPs were in coding regions, which led to synonymous variants and missense variants (Table 3). Notably, except for synonymous variants, 242 (19.0%) SNPs in coding regions resulted in variants of 43 protein coding genes (Supplementary Table S7). Of these genes, the *accD*, *clpP*, *ndhF*, *rpoC2*, *ycf1*, and *ycf2* harbor more than 10 missense SNPs. Notably, only four SNPs were identified between cultivated watermelon and its nearest wild species (*C. mucospermus*), and only one SNP (A/T at loci 78,734) was identified that altered the CDS of *petB*, which was validated by targeted pyrosequencing assays (Fig. 6), other three SNPs were in noncoding regions.

We also identified 234 small insertions or deletions (INDELs) among the chloroplast genomes of seven species. Most INDELs were in the intergenic regions and introns, and only 9 (3.8%) INDELs were located in six protein-coding regions of the *matK*, *rpl33*, *infA*, *rps19*, and *ycf1* genes (Supplementary Table S8). Notably, *matK*, *infA*, and *rps19* contain both variations, and a total of 45 protein-coding regions were altered by SNP or INDEL.

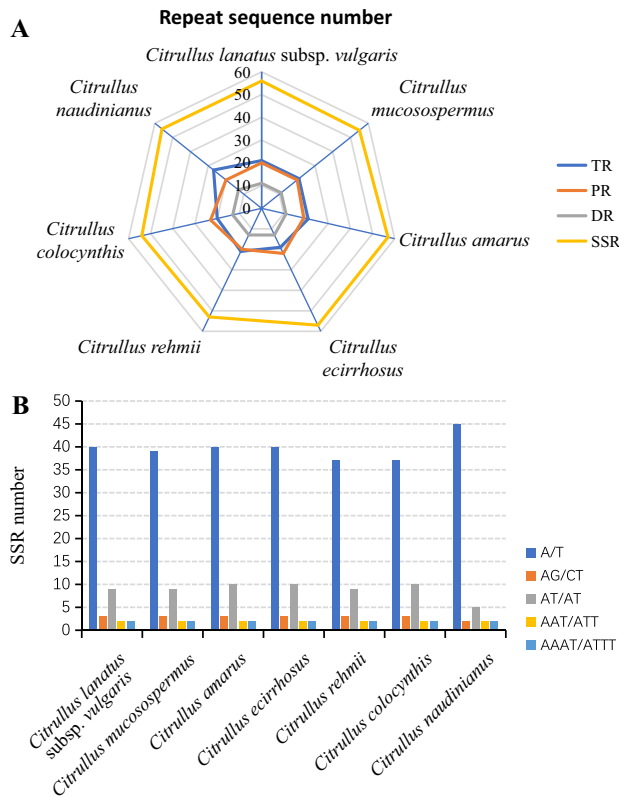


Figure 4. Different repeat sequence number showed in Radar-plot (A), Number of SSRs and types detected in the seven CPGs of *Citrullus* (B).

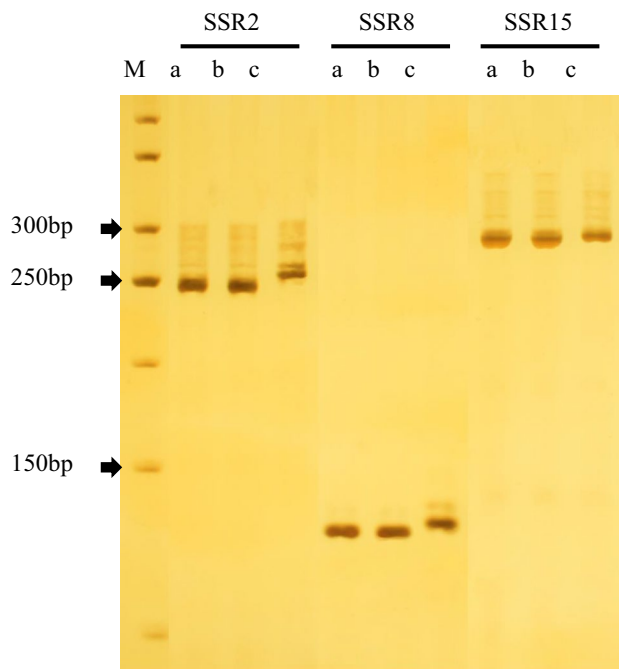


Figure 5. Three selected SSR makers were polymorphic among the *C. lanatus* subsp. *vulgaris* (W1-1), *C. mucosospermus* (PI186490) and *C. amarus* (PI269341), M: DL500 DNA marker, a: W1-1, b: PI186490, c: PI269341.

Type and region	Counts	Percent (%)
Upstream gene variant	761	59.9
Synonymous variant	260	20.5
Missense variant	242	19.0
Splice region variant and stop retained variant	1	0.1
Splice region variant and intron variant	3	0.2
Downstream gene variant	1	0.1
Stop gained	2	0.2
Stop lost and splice region variant	1	0.1

Table 3. The number of SNP effects with different type and region.

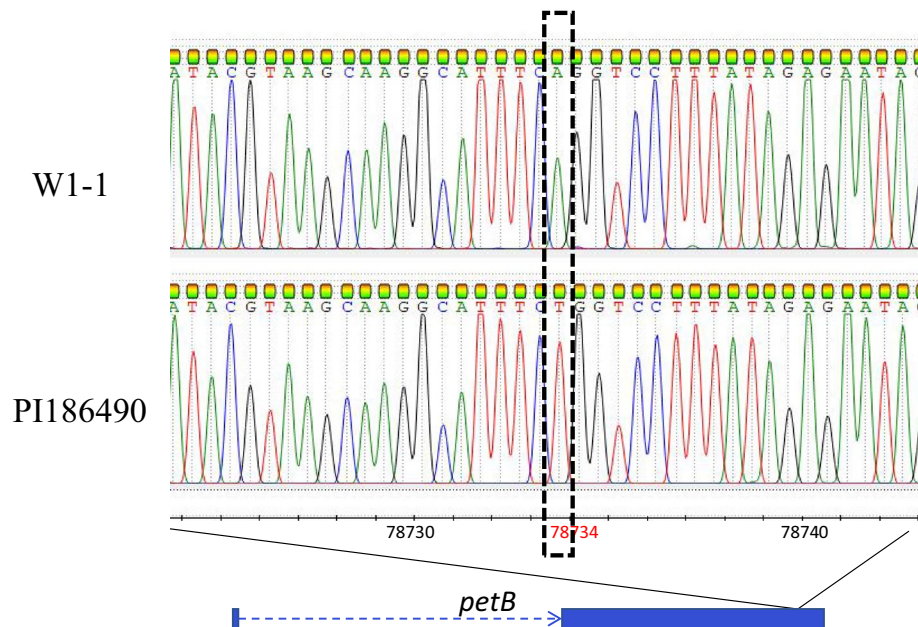


Figure 6. Targeted pyrosequencing assays of *petB* in W1-1 (*C. lanatus* subsp. *vulgaris*) and PI186490 (*C. mucospermus*).

Phylogenetic relationships analysis. The phylogenetic relationships of seven *Citrullus* species were analyzed with the ML and BI methods based on the CPG, CDS, SNP and DNA barcode, and the resulting topologies of the seven cucurbit plants were largely congruent, with highly supported values (Fig. 7 and Supplementary Fig. S5). Figure 7 shows the phylogenetic tree of CPG generated by the ML analysis, including two types of supporting values: the ML bootstrap (ML-BS), and BI posterior probability (BI-PP). All ML-BS and BI-PP values on the clade were 100 and 1.0, respectively. The phylogenetic tree revealed that the seven *Citrullus* species were grouped into a clade, and the *C. naudinianus* was at the base of the clade, followed by *C. colocynthis*, *C. rehmii*, *C. ecirrhosus*, *C. amarus*, *C. mucospermus*, and *C. lanatus* subsp. *vulgaris*.

Selective pressure events. The ratio (ω) was calculated to evaluate the selective pressure on the 79-consensus protein-encoding genes of the seven CPGs of *Citrullus*. Three genes (*atpB*, *clpP*, and *rpoC2*) were detected to be subject positive to selective pressure by EasyCodeML (Table 4). The ω in the M2a model (ω_2) changed from 177.5821 to 429.3408, and of the three genes subject to positive selective pressure, *clpP* had the greatest ω_2 value, which indicated that *clpP* could have undergone intense positive selection. Meanwhile, the amino acid sites in the three genes under positive selective pressure were determined, revealing that these genes harbor one intensively positive selective site (Table 4). Further analysis using the Ka/Ks calculator in Tbttools showed that the *clpP1* gene was positively selected ($\omega > 1$) between *C. lanatus* subsp. *vulgaris* and *C. naudinianus* (Supplementary Table S9).

Discussion

In our study, the CPGs of three species of the genus *Citrullus* were de novo assembled and annotated followed by a comparative analysis with four other *Citrullus* species, all methods were strictly conducted according to the above relevant experimental methods. The chloroplast genome size of the seven *Citrullus* species was $157,005 \pm 102$ bp,

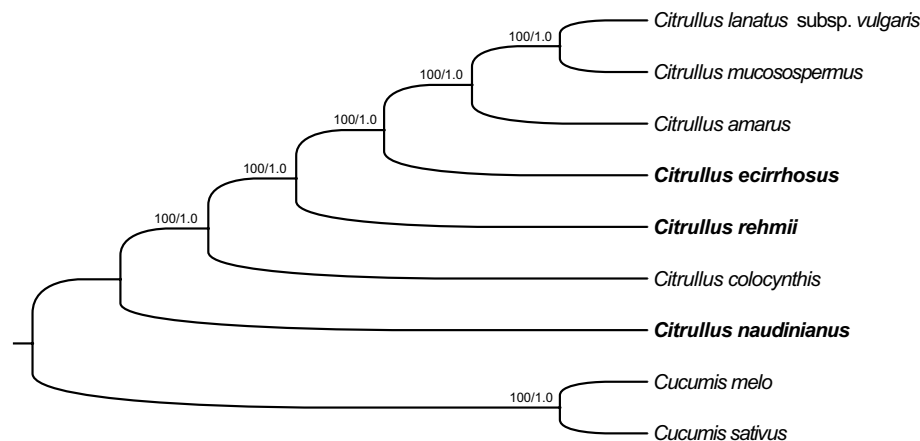


Figure 7. ML tree showing phylogenetic relationships among seven *Citrullus* species conducted using complete chloroplast sequences. *Cucumis sativus* and *Cucumis melo* acted as the out-group. The corresponding ML-BS/BI-PP values are given at each node. *C. ecirrhosus*, *C. rehmi*, and *C. naudinianus* were marked in bold characters.

Gene name	Model	np	LnL	ω 2(M2a)	LRTs(2 Δ LnL)	LRT <i>p</i> -value	Positive sites
<i>atpB</i>	M8	16	-2114.47	197.5187	38.81554	4.00E-09	52 Q*
	M7	14	-2133.88				
<i>clpP</i>	M8	16	-890.821	429.3408	34.30236	3.60E-09	108 S**
	M7	14	-907.973				
<i>rpoC2</i>	M8	16	-5901.01	177.5821	52.56322	0	937 G*
	M7	14	-5927.29				

Table 4. The summary of positive selective pressure analysis on genes in M2a, M7 vs. M8 model. *Means $p < 0.05$, **means $p < 0.01$.

suggesting that their genome size changed little over the course of evolution. The length change of the SC and IR regions are the dominant factor contributing to the differentiation of CPG size, the SSC regions of *Citrullus* have the highest variable coefficient among the SC and IR regions, and the correlation coefficient with CPG size was 0.87 (R^2) in our study, indicating that the variations of chloroplast genome size result from the expansion and contraction of SSC regions in *Citrullus*. The CPGs of *Citrullus* were composed of eight rRNA, 37 tRNA and 85–88 protein-encoding genes, and similar gene contents and types have been reported in other plants in Cucurbitaceae^{20–22}. The noncoding regions showed higher sequence divergence than the protein-encoding regions, suggesting that most sequence variation in the noncoding regions could be further exploited and utilized.

Repeated sequences could play a significant role in producing variation and rearranging genomes in CPG²³. The TR, PR, DR, and SSRs were analyzed in the CPGs of seven *Citrullus* species in our research, a lot of repeat sequences were located in the intergenic regions of the *Citrullus* CPGs, the resemble results to our study have been found in the CPGs of *Prunus* and *Quercus*¹⁰. The TR and mono SSR numbers in the chloroplast genome of *C. naudinianus* were greater than those in the other *Citrullus* species, which could contribute to the genetic divergence between *C. naudinianus* and other *Citrullus* species. Polymorphic SSRs in the chloroplast genome have been applied to study on the population genetics and evolutionary of *Citrullus*²⁴, in our study, Thirty-four SSRs were detected to be polymorphic among *Citrullus* species, suggesting our identified polymorphic SSRs could be developed to study on the genetics and evolutionary in *Citrullus*.

The SNPs and INDELS on the chloroplast genome have been used to study the diversity existed at the intra-species and inter-subspecies level^{25–28}. A total of 1271 SNPs were identified in the CPGs of seven *Citrullus* species, approximately 1 SNP/123 bp, which is less than that in the nuclear genome of 1 SNP/18 bp¹, suggesting that the mutation rate of the chloroplast genome is lower than that of the nuclear genome. Forty-three protein coding genes have been changed by 19.0% SNPs in the coding region, especially, the *accD*, *clpP*, *ndhF*, *rpoC2*, *ycf1*, and *ycf2* harbor more than 10 missense SNPs. These genes play a key role in plant photosynthesis and energy metabolism, but further research is needed to better understand the differences resulting from their mutations among *Citrullus* accessions. Although distinct differences in phenotype were observed between *C. lanatus* subsp. *vulgaris* and *C. mucospermus*, they have only four SNPs at their CPGs, further supporting that *C. mucospermus* is the most closely related wild species¹.

The high nucleotide diversity of noncoding regions and coding regions could be developed as molecular markers or DNA barcodes for species identification, taxonomy and phylogenetic relationship analysis²⁹. Six

intergenic regions (*trnS-trnG*, *rps16-trnQ*, *rps15-ycf1*, *rpl32-trnL*, and *psbE-petL*) with the highest nucleotide diversity (Pi) were identified in our study, while only *rps16-trnQ* could be used to develop specific DNA barcodes for the identification of germplasm resources in *Citrullus*. However, *rps16-trnQ* has not been used in the *Citrullus* taxonomy related study, but it has been widely used as a high-effect molecular marker and DNA barcode in other plant families^{28,30}. The *matk* and *rbcl* genes are commonly used as traditional DNA barcodes for Cucurbitaceae and other land plants^{31–33}, but their Pi values were 0.0021 and 0.0012, respectively, and most genes analyzed in this study have higher Pi values (Table S2), suggesting that *matk* and *rbcl* are not the best choice for using as DNA barcodes in *Citrullus*.

The CPG is an important molecular resource with better value for probing phylogenetic relationships than the whole nuclear genome in most plants¹⁰. Phylogenetic analysis based on the CPG has been used to reveal the phylogenetic relationships of species in Cucurbitaceae^{22,23,34}. Phylogenetic trees were built in our study based on the CPG, CDS, SNPs and DNA barcode (*rps16-trnQ*) and showed that *C. naudinianus* was sister to the other six species, followed by *Citrullus colocynthis*, *Citrullus rehmi*, *Citrullus ecirrhosus*, *Citrullus amarus*, *Citrullus mucososperms*, and *Citrullus lanatus* subsp. *vulgaris*, which was similar to the previously reported *Citrullus* phylogeny constructed based on phenotypic traits, ancient literature, murals, traditional chloroplast DNA barcode, and molecular markers of the nuclear genome^{1,2,35–39}, indicating that the phylogenetic tree based on our data has high reliability for resolving the relationship of the seven species, and the molecular markers (SNPs and DNA barcodes) identified in our study have potential to be used to identify and classify species in *Citrullus*. In addition, the completed chloroplast genome of *C. rehmi* (PI632755) was published in the NCBI genome (NC_035975) in a previous study (Unpublished), when *C. rehmi* (PI632755) was added to the phylogenetic analysis, the results showed that PI632755 is closer to *C. colocynthis* than to *C. rehmi* (Supplementary Fig. S6), suggesting that PI632755 could be classified as *C. colocynthis*, although it is still considered to be *C. rehmi* at U.S. National Plant Germplasm System and many previous studies^{1,40}.

Selective pressure analysis of genes could provide an important insight into the adaptive evolution of species. The KA/KS ratio is commonly applied to estimate the impacts of natural selection on creatures, and how the gene mutations promote the adaptive evolution of species^{22,41}. In the present study, three genes (*atpB*, *clpP1*, and *rpoC2*) that were subjected to positive selection. Among them, *atpB* encodes the ATP synthase CF1 beta subunit and plays an essential role in photosynthesis. The increasing requirements for light in watermelon inflicted strong selective pressures on genes involved in photosynthesis during plant adaptive evolution, which supports our finding of positive selection for *atpB*. The similar conclusions were also reported for other species in Cucurbitaceae^{22,34}. The product of the *clpP* gene is an ATP-dependent clp protease, which has been reported to participate in the transformation of chloroplast protein and shoot development^{12,22}. The fact that *clpP* underwent positive selection in our study could be related to the adaptive evolution of the vining feature like other species in Cucurbitaceae²². Similarly, *rpoC2* is validated to encode RNA polymerase β and could regulate the pollination process and sex differentiation in the genus *Siraitia* (Cucurbitaceae)²² and *sorghum*⁴². *Citrullus naudinianus* is dioecious, and the other species in *Citrullus* are monoecious¹, so the *rpoC2* mutation could be essential for the adaptive evolution of sex differentiation in *Citrullus*. Therefore, these chloroplast functional genes could not only participate in energy metabolism, but also in plant organ development and sex differentiation.

Conclusion

The present research reported the CPGs of *Citrullus rehmi*, *Citrullus naudinianus*, and *Citrullus ecirrhosus*, which were valuable resources for understanding the genus *Citrullus*. The CPG average length of *Citrullus* is 157,005, and all CPGs are composed of 130–133 genes. The CPG of the seven *Citrullus* species studied here showed a good level of similarity in terms of gene content, gene order and genome structure. However, the contraction and expansion of the SSC regions contribute to changes in the chloroplast genome size of *Citrullus*. The greatest sequence divergence was observed between *C. naudinianus* and *C. lanatus* subsp. *vulgaris* or *C. mucosospermus*, whereas the smallest sequence divergence was between *C. lanatus* subsp. *vulgaris* and *C. mucosospermus*. Thirty-four polymorphic SSRs, 1,271 SNPs and 234 INDELS were identified, and few variations in the coding regions led to the variation of 45 protein-coding genes. The *rps16-trnQ* can be developed as a high-resolution DNA barcode beneficial for taxonomy in *Citrullus*. Phylogenetic relationship analysis indicated that the clear phylogenetic relationship of *Citrullus* and molecular markers developed in our study could be applied to the species identification and classification of *Citrullus*. Three protein-coding genes (*atpB*, *clpP1*, and *rpoC2*) were subjected to selection pressure, which would promote the adaptation of *Citrullus* to the growing environment. Our research results provide a basis for understanding the genetic differences in the chloroplast genome of *Citrullus* and the domestication of cultivated watermelon.

Materials and methods

Plant material. Healthy fresh leaves were collected from young plants of *C. lanatus* subsp. *vulgaris* (W1-1), *C. mucosospermus* (PI186490), and *C. amarus* (PI269341) at the horticultural station of teaching and experimentation of Jiangxi Agricultural University (JXAU) in the spring of 2022. The total DNA of these plant materials were isolated from about 3 g of young leaves using a modified cetyltrimethylammonium bromide (CTAB) method. We used agarose gel electrophoresis (1%, 180 V for 45 min) and an Agilent BioAnalyzer 2100 (Agilent Technologies Inc., California, America) to evaluate the DNA quality and quantity. Approximately 300 μ g of total DNA from each sample was stored at -80 °C until use. Our experimental research and field studies on plants, including the collection of plant material, comply with the guidelines of JXAU and relevant institutional, national, and international guidelines and legislation.

Chloroplast genome assembly. To perform de novo assembly and compare the seven extant species in the genus *Citrullus*, for *C. rehmii* (SRR8751709), *C. ecirrhosus* (SRR8751725) and *C. naudinianus* (SRR8751817), we used Fastq-dump to download their clean sequence data with 150 bp paired-end reads of total DNA¹.

Next, the CPGs of the three *Citrullus* species above were assembled using the following method. First, to preclude the influence on chloroplast genome assembly from the consensus reads of nuclear and mitochondrial DNA with chloroplasts, 0.8-Gb sequence data were randomly drawn using the Sample function of Seqtk (v0.1) (<https://github.com/lh3/seqtk>) and then assembled by plasmidspades.py in SPAdes (v3.10.1)⁴³. Next, scaffolds with high coverage and length representing the chloroplast genome were extracted and ordered into a draft genome sequence by BLAST against the reference CPG of *C. lanatus* subsp. *vulgaris* (NC_014043.1). Gaps in the chloroplast draft genome sequence of each accession were repaired by GapCloser (v1.12-r6)⁴⁴ and overlap sequences at two flanks of gaps were deleted manually. Finally, the integrity and quality of the chloroplast genome sequence was then inspected and manually corrected by reference-guided mapping using BWA⁴⁵ and SAMtools⁴⁶ until the CPG was obtained.

Genome annotation. The three newly assembled CPGs were annotated using CpGAVAS2⁴⁷ and GeSeq⁴⁸, followed by manual adjustments of gene boundaries with errors through inspection using IGV v2.4⁴⁹. The final GenBank and Sequin format of the annotated information was produced in GB2sequin⁵⁰. All records with were then deposited in a Sequin format, and they can be viewed in the NCBI database. GenBank files were created using Sequin v1.5.0 (<http://www.ncbi.nlm.nih.gov/Sequin/index.html>) and then drew the chloroplast genome maps in OGDRAW v1.3.1⁵¹.

Comparative analysis of genome feature. To explore the intergenic variation among the CPGs of seven *Citrullus* species by genome comparison, the chloroplast reference genomes of *C. lanatus* subsp. *vulgaris*, *C. amarus*, and *C. colocynthis* that we had published in our past researches were also included^{14–16}, and the complete chloroplast genome of *C. mucosospermus* were directly obtained from the Organelle Genome Database (OGD; NCBI website). The data for the seven *Citrullus* species were listed in Table 5. The gene rearrangements at the boundary between the single copy (SC) region and inverted repeat (IR) region were inspected and analyzed by IRscope⁵² to understand the variation in binding regions within the CPGs among the seven *Citrullus* species. The chloroplast genome collinear analysis among the seven *Citrullus* species was conducted using Mauve Alignment in Geneious Prime⁵³.

Sequence divergence and nucleotide diversity analysis. To inspect the CPG sequence divergence of the seven *Citrullus* species, the mVISTA web tool was used to align and visualize the results with default parameters⁵⁴. The CPG of *C. lanatus* subsp. *vulgaris* was used as reference, and then the pairwise distances based on the CPG of seven *Citrullus* species was estimated by MAFFT v7⁵⁵ and MEGA X⁵⁶. Nucleotide diversity in the CPGs of the seven *Citrullus* species were scanned using DNA polymorphism in DNAsp v6.12.03 with an 800-bp window and 200-bp step size⁵⁷. Parsimony-informative base sites (PIC), variable sites and K-2P distance of selected DNA barcodes were analyzed in MEGA v11 software⁵⁸.

Repeated sequence identification. Dispersed repeats (DR), palindromic repeats (PR), long tandem repeats (TR) (greater than 7 bp), and simple sequence repeats (SSR) were explored in the CPGs of seven *Citrullus* species. Vmatch software was applied to identify the DR and PR with a minimal repeat size of 30 bp, a hamming distance is 3, and a similarity percentage between two repeat copies greater than 90%, the overlapping repeats in the results were merged into one repeat motif by manual correction. TR at least 7 bp in length was detected by Tandem Repeats Finder program with the alignment parameters for match, mismatch, and INDELs set at 2, 7, and 7, respectively⁵⁹. SSRs were determined by MISA with thresholds of 10, 5, 4, 4, and 4 for mono-, di-, tri-, tetra-, penta-, and hexa-nucleotides, respectively⁶⁰.

Variation calling. To identify the single nucleotide polymorphisms (SNPs) and insertions or deletions (INDELs) of the CPGs between cultivated watermelon (*C. lanatus* subsp. *vulgaris*) and six other *Citrullus* species, we downloaded their total DNA sequence data from the NCBI SRA database, as shown in Table 5. Then,

Species	Accession	Chloroplast genome ^a	DNA sequence data accession ^b
<i>C. rehmii</i>	Grif16376	–	SRR8751709
<i>C. ecirrhosus</i>	Grif16945	–	SRR8751725
<i>C. naudinianus</i>	PI 596,694	–	SRR8751817
<i>C. lanatus</i> subsp. <i>vulgaris</i>	W1-1	NC_032008.1	–
<i>C. colocynthis</i>	PI 374,216	NC_035727.1	SRR8751849
<i>C. amarus</i>	Grif16135	NC_035974.1	SRR8751536
<i>C. mucosospermus</i>	PI 249,010	NC_033899.1	SRR8751721

Table 5. A list of the seven *Citrullus* species used in the study. ^aPlant chloroplast genome directly download from the NCBI ORG database. ^bPlant whole genome re-sequence data downloaded from the NCBI SRA database according to the accession in Guo et al.¹

Primer name	Forward primer sequence	Reverse primer sequence	Tm (°C)	Product size (bp)
petB_p	TTCAGGCGATTGCAGATGAT	TTTCGTGGAGTTGAAGGAAA	57	742
SSR2	GTACGGCTCGAGGAAATCA	GGGGTTAGAGACCGCTCAAT	60	238
SSR8	ATGCATTCAGATGAAAAGGC	CCGAACGTGAAACTTTGGTT	60	126
SSR15	GGTTATGCCACGATGTCCTT	GACATGTGGCTTGCTGGTA	60	280

Table 6. Designed primer for SNP validation and molecular markers.

their clean paired reads were aligned to the CPG of *C. lanatus* subsp. *vulgaris* using BWA-mem, and the aligned results were sorted and indexed by SAMtools⁴⁵. Finally, the SNP and INDEL were called by VarScan⁶¹. Variants with quality scores below 20 and homopolymers were filtered, and the effect of remaining variants in variant call format (VCF) format were annotated and predicted by using SnpEff⁶² based on the transcript definitions of the chloroplast reference genome of *C. lanatus* subsp. *vulgaris*. The high severity effects in the report were chosen to represent the effect of variants.

Phylogenetic analyses. The phylogenetic tree was built based on the CPG, CDS, SNPs and hypervariable fragments using the maximum likelihood (ML) and Bayesian inference (BI) approaches. The CPG of *Cucumis sativus* and *Cucumis melo* served as the out-group. ML was implemented using PhyML⁶³ with a GTR + G model; to assess node support, a bootstrap analysis was performed based on 1,000 replicates in UGENE v1.29.0⁶⁴. The BI analysis was conducted in MrBayes v3.2.6⁶⁵, for which the Markov chain Monte Carlo (MCMC) method was run for 1 million generations. Trees were sampled and diagnosed every 1,000 generations, with the first 25% of samples discarded as burn-in. The remaining trees were used to build a 50% majority-rule consensus tree. A steady state was reached if the average standard deviation of the split frequencies was < 0.01. The final phylogenetic tree was edited and visualized in FigTree v1.4.2 software (<http://tree.bio.ed.ac.uk/software/figtree/>).

Selective pressure analysis. Protein-encoding genes could be subjected to selective pressure, thus, all consensus protein-encoding genes of the CPGs of the seven *Citrullus* species included in this study were analyzed for selective pressure using Easy-CodeML⁶⁶ and TTools⁶⁷. In Easy-CodeML with the site model, the non-synonymous (Ka) and synonymous (Ks) substitution ratios and likelihood ratio test (LRT) values were calculated. The Ka/Ks ratio (ω) and the LRT values were integrated to explore the selective pressure on amino acid sites using naive empirical Bayes and Bayes empirical Bayes methods in the M7 versus M8 model⁶⁶. Meanwhile, in order to determine the selective pressure between *C. lanatus* subsp. *vulgaris* and the other six *Citrullus* species, the Ka/Ks ratios were calculated by the Ka/Ks calculator of TTools according to their paired protein-encoding genes.

SNP validation and SSR marker development. To validate the missense SNP of *petB* between cultivated watermelon and *C. mucospermus*, the *petB* gene was cloned from W1-1 (*C. lanatus* subsp. *vulgaris*) and PI186490 (*C. mucospermus*). First, the primer was designed according to the CDS of *petB* in W1-1 (Table 6). Then, total mRNA was extracted from leaves and reverse transcribed for first-strand cDNA synthesis using the M5 Sprint qPCR RT kit according to the manufacturer's instructions (Mei5 Biotechnology, Co, Ltd). The DNA fragments were amplified and purified according to the operation instructions of the Hieff Canace Plus PCR Master Mix and TIAgel purification kit, and sequenced using the targeted pyrosequencing assay in Sangon Biotech (Shanghai) Co., Ltd. Three selected SSRs with flanking sequences were used to design primers in Primer Premier 5.0⁶⁸ software with default parameters (Table 6). The polymorphism of molecular markers was verified among W1-1 (*C. lanatus* subsp. *vulgaris*), PI186490 (*C. mucospermus*), and PI269341 (*C. amarus*) using PCR amplification, and PCR assay was conducted according to our previous study⁶⁹. The PCR products were separated with by 6% denaturing polyacrylamide gel electrophoresis (PAGE) and detected by silver staining.

Data availability

The re-sequencing data of Grif16376 (*C. rehmii*), Grif16945 (*C. ecirrhosus*) and PI596694 (*C. naudinianus*) were in GenBank (<https://www.ncbi.nlm.nih.gov/>) with accession number SRR8751709, SRR8751725 and SRR8751817, respectively. Fully assembled and annotated plastomes also have been deposited in GenBank (*C. rehmii*: ON597627, *C. ecirrhosus*: MZ577999, *C. naudinianus*: MZ577998).

Received: 1 January 2023; Accepted: 23 April 2023

Published online: 25 April 2023

References

- Guo, S. *et al.* Resequencing of 414 cultivated and wild watermelon accessions identifies selection for fruit quality traits. *Nat. Genet.* **51**, 1616–1623. <https://doi.org/10.1038/s41588-019-0518-4> (2019).
- Renner, S. S. *et al.* A chromosome-level genome of a Kordofan melon illuminates the origin of domesticated watermelons. *Proc. Natl. Acad. Sci.* **118**, e2101486118. <https://doi.org/10.1073/pnas.2101486118> (2021).
- Guo, S. *et al.* The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. *Nat. Genet.* **45**, 51–58. <https://doi.org/10.1038/ng.2470> (2013).
- Renner, S. S., Sousa, A. & Chomicki, G. Chromosome numbers, Sudanese wild forms, and classification of the watermelon genus *Citrullus*, with 50 names allocated to seven biological species. *Taxon* **66**, 1393–1405. <https://doi.org/10.12705/666.7> (2017).
- Huseini, H. F. *et al.* The Clinical Investigation of *Citrullus colocynthis* (L.) Schrad Fruit in treatment of Type II Diabetic Patients: A randomized, double blind, placebo-controlled clinical trial. *Phytother. Res.* **23**, 1186–1189. <https://doi.org/10.1002/ptr.2754> (2009).

6. Ayyad, S.-E.N. *et al.* In vitro and in vivo study of cucurbitacins-type triterpene glucoside from *Citrullus colocynthis* growing in Saudi Arabia against hepatocellular carcinoma. *Environ. Toxicol. Pharmacol.* **33**, 245–251. <https://doi.org/10.1016/j.etap.2011.12.010> (2012).
7. Hussain, A. I. *et al.* *Citrullus colocynthis* (L.) Schrad (bitter apple fruit): A review of its phytochemistry, pharmacology, traditional uses and nutritional potential. *J. Ethnopharmacol.* **155**, 54–66. <https://doi.org/10.1016/j.jep.2014.06.011> (2014).
8. Chomicki, G., Schaefer, H. & Renner, S. S. Origin and domestication of Cucurbitaceae crops: Insights from phylogenies, genomics and archaeology. *New Phytol.* <https://doi.org/10.1111/nph.16015> (2019).
9. Davis, A. R., Perkins-Veazie, P., Collins, J. & Levi, A. LSW-177 and LSW-194: Red-fleshed watermelon lines with low-total soluble solids. *HortScience* **43**, 538–539 (2008).
10. Xue, S. *et al.* Comparative analysis of the complete chloroplast genome among *Prunus mume*, *P. armeniaca*, and *P. salicina*. *Hortic. Res.* **6**, 89. <https://doi.org/10.1038/s41438-019-0171-1> (2019).
11. Kuroda, H. & Maliga, P. The plastid *clpP1* protease gene is essential for plant development. *Nature* **425**, 86–89. <https://doi.org/10.1038/nature01909> (2003).
12. Gao, Y. *et al.* Chloroplast translational regulation uncovers nonessential photosynthesis genes as key players in plant cold acclimation. *Plant Cell* **34**, 2056–2079. <https://doi.org/10.1093/plcell/coac056%JThePlantCell> (2022).
13. Hollingsworth, P. M. Refining the DNA barcode for land plants. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 19451–19452. <https://doi.org/10.1073/pnas.1116812108> (2011).
14. Cui, H., Ding, Z., Zhu, Q., Wu, Y. & Gao, P. Population structure and genetic diversity of watermelon (*Citrullus lanatus*) based on SNP of chloroplast genome. *3 Biotech* <https://doi.org/10.1007/s13205-020-02372-5> (2020).
15. Zhu, Q. *et al.* The complete chloroplast genome sequence of the *Citrullus lanatus* L. Subsp. *Vulgaris* (Cucurbitaceae). *Mitochondrial DNA B Resour.* **1**, 943–944. <https://doi.org/10.1080/23802359.2016.1261611> (2016).
16. Zhu, Q. *et al.* The complete chloroplast genome sequence of the *Citrullus colocynthis* L. (Cucurbitaceae). *Mitochondrial DNA B Resour.* **2**, 480–482. <https://doi.org/10.1080/23802359.2017.1361351> (2017).
17. Shi, C., Wang, S., Zhao, F., Peng, H. & Xiang, C. L. Full chloroplast genome assembly of 11 diverse watermelon accessions. *Front. Genet.* **8**, 46. <https://doi.org/10.3389/fgene.2017.00046> (2017).
18. Sapienza, C., Rose, M. R. & Doolittle, W. F. High-frequency genomic rearrangements involving archaeobacterial repeat sequence elements. *Nature* **299**, 182–185. <https://doi.org/10.1038/299182a0> (1982).
19. Curci, P. L., De Paola, D., Danzi, D., Vendramin, G. G. & Sonnante, G. Complete chloroplast genome of the multifunctional crop globe artichoke and comparison with other Asteraceae. *PLoS ONE* **10**, e0120589. <https://doi.org/10.1371/journal.pone.0120589> (2015).
20. Rodriguez-Moreno, L. *et al.* 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> (2011).
21. Pläder, W., Yukawa, Y., Sugiura, M. & Maleszky, S. The complete structure of the cucumber (*Cucumis sativus* L.) chloroplast genome: Its composition and comparative analysis. *Cell. Mol. Biol. Lett.* **12**, 584–594 (2007).
22. Shi, H. *et al.* Complete chloroplast genomes of two *Siraitia merrillii* species: Comparative analysis, positive selection and novel molecular marker development. *PLoS ONE* **14**, e0226865. <https://doi.org/10.1371/journal.pone.0226865> (2019).
23. Zhang, X. *et al.* Completion of eight *Gynostemma* BL. (Cucurbitaceae) Chloroplast genomes: characterization, comparative analysis, and phylogenetic relationships. *Front. Plant Sci.* **8**, 1583. <https://doi.org/10.3389/fpls.2017.01583> (2017).
24. Dane, F., Lang, P. & Bakhtiyarova, R. Comparative analysis of chloroplast DNA variability in wild and cultivated *Citrullus* species. *Theor. Appl. Genet.* **108**, 958–966. <https://doi.org/10.1007/s00122-003-1512-9> (2004).
25. Cui, H., Ding, Z., Zhu, Q., Wu, Y. & Gao, P. Population structure and genetic diversity of watermelon (*Citrullus lanatus*) based on SNP of chloroplast genome. *3 Biotech* **10**, 374. <https://doi.org/10.1007/s13205-020-02372-5> (2020).
26. Zhang, Y. *et al.* Chloroplast subspecies-specific SNP detection and its maternal inheritance in *Brassica oleracea* L. by Using a dCAPS Marker. *J. Hered.* **103**, 606–611. <https://doi.org/10.1093/jhered/ess006> (2012).
27. Umali, R. & Nakamura, I. Identification of dCAPS markers that discriminate A and B cytoplasm in Banana (*Musa* spp.). *Plant Biotechnol.* **20**, 159–164. <https://doi.org/10.5511/plantbiotechnology.20.159> (2003).
28. Sheng, J. *et al.* The complete chloroplast genome sequences of five *Miscanthus* species, and comparative analyses with other grass plastomes. *Ind. Crops Prod.* **162**, 113248. <https://doi.org/10.1016/j.indcrop.2021.113248> (2021).
29. Song, Y. *et al.* Development of chloroplast genomic resources for *Oryza* species discrimination. *Front. Plant Sci.* <https://doi.org/10.3389/fpls.2017.01854> (2017).
30. Bello, A., Stirton, C. H., Chimphango, S. B. & Muasya, A. M. Phylogenetic relationships and biogeography of Psoraleae (Fabaceae). *Bot. J. Linn. Soc.* **200**, 39–74. <https://doi.org/10.1093/botlinn/boac002%JBotanicalJournaloftheLinneanSociety> (2022).
31. Group, C. P. W. A DNA barcode for land plants. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 12794. <https://doi.org/10.1073/pnas.0905845106> (2009).
32. De Vere, N., Rich, T. C. G., Trinder, S. A. & Long, C. DNA barcoding for plants. *Methods Mol. Biol.* **1245**, 101–118. https://doi.org/10.1007/978-1-4939-1966-6_8 (2015).
33. Sebastian, P., Schaefer, H., Telford, I. R. & Renner, S. S. Cucumber (*Cucumis sativus*) and melon (*C. melo*) have numerous wild relatives in Asia and Australia, and the sister species of melon is from Australia. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 14269–14273. <https://doi.org/10.1073/pnas.1005338107> (2010).
34. Zhang, X. *et al.* Comparative analyses of chloroplast genomes of Cucurbitaceae species: Lights into selective pressures and phylogenetic relationships. *Molecules* **23**, 2165. <https://doi.org/10.3390/molecules23092165> (2018).
35. Paris, H. S. Origin and emergence of the sweet dessert watermelon, *Citrullus lanatus*. *Ann. Bot.* **116**, 133–148. <https://doi.org/10.1093/aob/mcv077> (2015).
36. Paris, H. S., Daunay, M. C. & Janick, J. Medieval iconography of watermelons in Mediterranean Europe. *Ann. Bot.* **112**, 867–879. <https://doi.org/10.1093/aob/mct151> (2013).
37. Dane, F. & Lang, P. Sequence variation at cpDNA regions of watermelon and related wild species: Implications for the evolution of *Citrullus* haplotypes. *Am. J. Bot.* **91**, 1922–1929. <https://doi.org/10.3732/ajb.91.11.1922> (2004).
38. Zhang, H. *et al.* Identification and validation of a core set of microsatellite markers for genetic diversity analysis in watermelon, *Citrullus lanatus* Thunb. Matsum. & Nakai. *Euphytica* **186**, 329–342. <https://doi.org/10.1007/s10681-011-0574-z> (2011).
39. Chomicki, G. & Renner, S. S. Watermelon origin solved with molecular phylogenetics including Linnaean material: Another example of museomics. *New Phytol.* **205**, 526–532. <https://doi.org/10.1111/nph.13163> (2015).
40. Tetteh, A. Y., Wehner, T. C. & Davis, A. R. Identifying resistance to powdery mildew race 2W in the USDA-ARS watermelon germplasm collection. *Crop Sci.* **50**, 933–939. <https://doi.org/10.2135/cropsci2009.03.0135> (2010).
41. Zhao, K. *et al.* Comparative analyses of chloroplast genomes from 14 *Zanthoxylum* species: Identification of variable DNA markers and phylogenetic relationships within the Genus. *Front Plant Sci* **11**, 605793. <https://doi.org/10.3389/fpls.2020.605793> (2021).
42. Chen, Z., Schertz, K. F., Mullet, J. E., DuBell, A. & Hart, G. E. Characterization and expression of rpoC2 in CMS and fertile lines of sorghum. *Plant Mol. Biol.* **28**, 799–809. <https://doi.org/10.1007/bf00042066> (1995).
43. Bankevich, A. *et al.* SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477. <https://doi.org/10.1089/cmb.2012.0021> (2012).

44. Luo, R. *et al.* SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**, 18. <https://doi.org/10.1186/2047-217X-1-18> (2012).
45. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324> (2009).
46. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352> (2009).
47. Shi, L. *et al.* CPGAVAS2, an integrated plastome sequence annotator and analyzer. *Nucl. Acids Res.* **47**, W65–W73. <https://doi.org/10.1093/nar/gkz345> (2019).
48. Tillich, M. *et al.* GeSeq - versatile and accurate annotation of organelle genomes. *Nucl. Acids Res.* **45**, w6–w11. <https://doi.org/10.1093/nar/gkx391> (2017).
49. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26. <https://doi.org/10.1038/nbt.1754> (2011).
50. Lehwark, P. & Greiner, S. 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> (2019).
51. Greiner, S., Lehwark, P. & Bock, R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: Expanded toolkit for the graphical visualization of organellar genomes. *Nucl. Acids Res.* **47**, W59–W64. <https://doi.org/10.1093/nar/gkz238> (2019).
52. Amiryousefi, A., Hyvönen, J. & Poczai, P. IRscope: an online program to visualize the junction sites of chloroplast genomes. *Bioinformatics* **34**, 3030–3031. <https://doi.org/10.1093/bioinformatics/bty220%JBioinformatics> (2018).
53. Darling, A. E., Mau, B. & Perna, N. T. progressiveMauve: Multiple genome alignment with gene gain. *Loss and Rearrangement. PloS one* **5**, e11147. <https://doi.org/10.1371/journal.pone.0011147> (2010).
54. Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M. & Dubchak, I. VISTA: computational tools for comparative genomics. *Nucl. Acids Res.* **32**, W273 (2004).
55. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780. <https://doi.org/10.1093/molbev/mst010> (2013).
56. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).
57. Rozas, J. *et al.* DnaSP 6: DNA sequence polymorphism analysis of large datasets. *Mol. Biol. Evol.* **34**(12), 3299–3302 (2017).
58. Tamura, K., Stecher, G. & Kumar, S. MEGA11: Molecular evolutionary genetics analysis version 11. *Mol. Biol. Evol.* **38**, 3022–3027. <https://doi.org/10.1093/molbev/msab120> (2021).
59. Benson, G. Tandem repeats finder: A program to analyze DNA sequences. *Nucl. Acids Res.* **27**, 573–580 (1999).
60. Beier, S., Thiel, T., Munch, T., Scholz, U. & Mascher, M. MISA-web: a web server for microsatellite prediction. *Bioinformatics* **33**, 2583–2585. <https://doi.org/10.1093/bioinformatics/btx198> (2017).
61. Koboldt, D. C. *et al.* VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576. <https://doi.org/10.1101/gr.129684.111> (2012).
62. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80–92. <https://doi.org/10.4161/fly.19695> (2012).
63. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst Biol* **59**, 307–321. <https://doi.org/10.1093/sysbio/syq010> (2010).
64. Okonechnikov, K., Golosova, O., Fursov, M., team, U. Unipro UGENE: A unified bioinformatics toolkit. *Bioinformatics* **28**, 1166–1167. <https://doi.org/10.1093/bioinformatics/bts091> (2012).
65. Ronquist, F. *et al.* MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542. <https://doi.org/10.1093/sysbio/sys029> (2012).
66. Gao, F. *et al.* EasyCodeML: A visual tool for analysis of selection using CodeML. *Ecol. Evol.* <https://doi.org/10.1002/ece3.5015> (2019).
67. Chen, C. *et al.* TBtools: An integrative toolkit developed for interactive analyses of big biological data. *Mol Plant* **13**, 1194–1202. <https://doi.org/10.1016/j.molp.2020.06.009> (2020).
68. Lalitha, S. Primer Premier 5. *Biotech Softw. Int. Rep.* **1**, 270–272. <https://doi.org/10.1089/152791600459894> (2000).
69. Liu, S. *et al.* Mapping of quantitative trait loci for lycopene content and fruit traits in *Citrullus lanatus*. *Euphytica* **202**, 411–426 (2015).

Acknowledgements

The authors thank for publishing the whole genome sequence data used in this study.

Author contributions

Q.L.Z. designed the experiment and acquired fundings to support the study; R.Z. and W.H. performed the experiments; Q.Z. did all the figures; C.Z. and P.W. analyzed the experimental data and wrote the draft manuscript; L.S., F.L. and S.L. reviewed and edited the draft manuscript.

Funding

This work was supported by grants from the National Nature Science Foundation of China (31960607), the Science Foundation of Jiangxi Province project (20202ACBL215006), the College Students' Innovation and Entrepreneurship Training Program under Grant number (S202210410038).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-023-34046-6>.

Correspondence and requests for materials should be addressed to Q.L.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023