Complete chloroplast genome of a Peruvian landrace of Cucurbita moschata, loche, and its comparative analysis with other relative species

Carla L. Saldaña
Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas (UNTRM)
Richard Estrada
Instituto Nacional de Innovación Agraria (INIA)
Esther Suca
Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas (UNTRM)
Camila Cruz
Instituto Nacional de Innovación Agraria (INIA)
Edgardo Vilcara
Instituto Nacional de Innovación Agraria (INIA)
Carlos I. Arbizu
carlos.arbizu@untrm.edu.pe
Instituto Nacional de Innovación Agraria (INIA)

Research Article

Keywords: Cucurbitaceae, genomics, Lambayeque, genetic resources, genome, next-generation sequencing

Posted Date: October 10th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-5034257/v1

Additional Declarations: No competing interests reported.

Abstract

Background Loche (*Cucurbita moschata*) is a pre-Columbian squash that is cultivated exclusively in the northern coast of Peru, Lambayeque. This crop is notable for the presence of warts in its skin and for its vegetative reproduction as it commonly lacks of seeds in fruits. Currently, loche may be considered a new product for international markets, recognizing the gastronomy of Lambayeque in the world and dynamizing the Peruvian agro-export area. However, genomic information about this squash is very limited.

Results In this study, the complete chloroplast (cp) genome of loche from Lambayeque was sequenced and annotated. Clean reads were obtained by PE 150 library and the Illumina HiSeq 2500 platform. The complete chloroplast (cp) genome of *C. moschata* has a 157,592 bp in length with typical quadripartite structure, containing a large single copy (LSC) region (88,192 bp) and 18,181 bp small single-copy (SSC) region, separated by two inverted repeat (IR) regions (25,613 bp). The annotation of *C. moschata* cp genome predicted 81 protein-coding genes (CDS), 8 ribosomal RNA (rRNA) genes, 38 transfer RNA (tRNA) genes and 01 pseudogen. A total of 59 simple sequence repeats (SSR) of this cp genome were divided into mononucleotide (43), dinucleotide (7), trinucleotide (2), tetranucleotide (6), and hexanucleotide (1). The highest percentage of identity was observed for *C. moschata* and *C. maxima* (0.99) while the lowest for *C. moschata* and *Cucumis sativus* (0.68). *Cucurbita pepo* is the closest relative to loche.

Conclusions The cp genome of loche is similar to other cucurbit species and possesses 127 genes in total. Moreover, a total of 59 SSR were identified in this cp genome. A higher percentage of identity is evidenced between *C. moschata* and *C. maxima* whereas higher divergence values with *Cucumis sativus*. This study reports for the first time the genome organization, gene content, and structural features of the chloroplast genome of a Peruvian squash landrace, that is commonly cultivated in a restricted area in northern Peru, providing valuable information for genetic and evolutionary studies in the genus *Cucurbita*.

Background

Cucurbitaceae are among the most economically, nutritionally important and large plant families (100 genera and 1000 species) [1][2]. This family includes cucumber, melon, watermelon, pumpkins and squashes [3]. *Cucurbita* genus contains earliest domesticated plant species [4], like *C. moschata* Duchesne, *C. maxima* Duchesne and *C. pepo* L. These species are appreciated due to their nutritional and medical properties as they are able to reduce the risk of coronary heart disease, as well as blood glucose and serum cholesterol levels [5]. This is due to the presence of phytoconstituents such as tannins, flavonoids, and terpenoid cucurbitacins, which provide them with antioxidant, anti-inflammatory, antidiabetic, and other properties [6]. Although the primary and most consumed part of the plant is the mature fruit, the leaves, flowers, and shoots are also used for culinary purposes [7]. Additionally, the seeds are an important source of protein, essential fatty acids, and linoleic acid [8]. Moreover, plants belonging to the Cucurbitaceae family are used as a nutritional supplement in livestock, poultry, and aquaculture [9].

Cucurbita moschata is an important vegetable in America. It was cultivated since the pre-Colombian era and probably was domesticated in Mexico and South America independently [10]. Currently, it constitutes an important part of the traditional polyculture systems in Mexico and Peru together with corn, beans, among others. In Peru there is a squash landrace called “loche”. This landrace has been cultivated at least since the Chimú culture, and currently is grown exclusively and traditionally in the north coast of Peru, in the geographical departments of Lambayeque and is unknown elsewhere (Andres et al., 2006). Interestingly, this squash is vegetatively propagated, possesses low genetic diversity, and mostly lacks of seeds in fruits and also presents warts in its skin [12]. In addition, this crop represents an important component in the northern Peruvian gastronomy, and is also part of the economy and culture of northern Peru.

Chloroplast plays an import role into vital metabolic events, including photosynthesis, lipid synthesis and amino acid. Angiosperm plastid genomes exist in linear form and most commonly circular form [13]. Its size varies between 120 ± 150

kb, with a quadripartite structure containing a large and small single copy regions (LSC and SSC) separated by two inverted repeats regions (IRA and IRB). The chloroplast includes 110 ± 130 genes involved in photosynthesis, translation and transcription [14]. Due to conserved gene, content, organization and order makes them well suited for evolutionary studies [15], phylogenetic analyses [16], population structure [17], structural rearrangements, pseudogenes or additional mutation events [15] and genetic engineering studies [18]. Currently, most cp genomes of important Cucurbitaceae species were sequenced. Cong et.al [19] revealed that *C. ficifolia* possesses 157,533 bp in length, a pair of inverted repeats regions (IRs) of 25,639 bp, separated by large single copy (LSC) (88,112 bp) and small single copy (SSC) (18,143 bp). This cp genome encodes 86 protein-coding genes, eight rRNA genes and 36 tRNA genes. Finally, maximum likelihood phylogenetic analysis revealed that *C. ficifolia* is a base clade of genus *Cucurbita* and closer to *C. maxima*. *Cucumis melo* 'Shengkaihua' is another important species whose cp genome has already been revealed. Using next-generation sequencing, the entire cp genome was obtained, possessing 156,017 bp in length with typical structure: LSC (86,335 bp), SSC (18,088 bp), separated by a pair of 25,797 bp (IR regions). This genome contained 133 genes, including 88 protein-coding genes, 37 tRNA genes, and eight rRNA genes. The GC content of the genome is 36.9%. The phylogenetic tree reconstructed by 24 chloroplast genomes revealed that *C. melo* is most related to *Cucumis melo* var. *inodorus*.

The increased use of next-generation sequencing technologies has allowed access to a large amount of nucleotide data, facilitating comparative studies to better understand phylogenetic hypotheses [20]. Zhang et al. [21] pointed out that Cucurbitaceae is the fourth most important economic plant family, mainly distributed in tropical and subtropical regions. They compared and described the complete cp genome sequences of ten representative species from Cucurbitaceae. The cp genomes of the ten species ranged from 155,293 bp (*C. sativus*) to 158,844 bp (*M. charantia*). Phylogenetic analysis strongly supported the position of *Gomphogyne, Hemsleya*, and *Gynostemma* as the relatively original lineage in Cucurbitaceae. On the other hand, [22] carried out a comparative analysis of the cp genome of nine varieties of *Cucumis melo*, which represented the morphological diversity of two subspecies, *Cucumis melo* ssp. m*elo* and *C. melo* ssp. *agrestis*. This study demonstrated that the cp genome of melon is relatively conserved, and the phylogenetic results indicated that ssp. *melo* and ssp. *agrestis* formed a monophyletic group, providing a quick and simple method to identify and differentiate them. Therefore, large-scale comparisons of cp sequences are of great interest, as they provide solid evidence for taxonomic studies, species identification, and understanding the mechanisms underlying evolution in Cucurbitaceae.

To date, *C. moschata* is still considered a neglected crop [23] because the agricultural, economic and biological importance of this cucurbit is still unknown. Here we sequenced for the first time the complete cp genome of this Peruvian orphan crop and compared it with other eight important species within the Cucurbitaceae family. This work adds valuable information to the complete chloroplast genomics of Cucurbitaceae, providing a solid foundation for the development of DNA barcoding at the species level, the use of microsatellites (SSRs) as polymorphic molecular markers, as well as for studies on the evolution and molecular identification of *C. moschata* cultivars.

Results

C. moschata chloroplast genome organization and features

Complete cp sequence of *C. moschata* exhibits 157,592 bp in size and has a quadripartite structural organization containing a large single copy (LSC, 88,192 bp), a pair of inverted repeats (IRa and IRb 25,613 bp each) and small single copy (SSC, 18,181 bp) (Table 1 and Fig. 1). The percentage of GC of the IR region was 43.03%, which is higher than that of LSC and SSC regions with 35.90% and 32.02% respectively (Table 1). In the *C. moschata* cp genome, 127 genes were predicted in total; 81 were protein-coding genes, 38 transfer RNA (tRNAs) genes, eight ribosomal RNA (rRNAs) genes, and one pseudogene (Table 2). Of these, six protein coding genes, four rRNAs, and eight tRNAs are duplicated in the IR regions. A total of ten protein-coding genes and eight tRNAs genes contained a single intron, whereas three genes exhibited two

introns each. The *rpsl2* gene was predicted to be trans-spliced with its 5-end located at the LSC region and the 3- ends with a copy located in each of the two IR regions.

The cp genome of “loche” possesses in its IR region 18 duplicated genes: five protein-coding genes such as *rpl2*, *rpl23*, *rps7*, *rps12*, and *ndhB*; seven tRNAs as *trnI-CAU*, *trnL-CAA*, *trnV-GAC*, *trnI-GAU*, *trnA-UGC*, *trnR-AGC*, and *trnN-GUU*; four rRNAs genes as *rrn23*, *rrn16*, *rrn5*, *rrn4.5* (Fig. 1), and the 5´-end of *ycf1* gene. The SSC region contained 12 protein-coding and one tRNA gene, whereas. LSC region contained 69 protein-coding and 22 tRNAs.

Table 1
Chloroplast genomes of “loche” (*C. moschata*) and seven Cucurbitaceae species

Genome Characteristics	*Citrullus lanatus*	*Cucumis melo*	*Cucumis sativus*	*Cucurbita ficifolia*	*Cucurbita maxima*	*Cucurbita moschata*	*Cucurbita pepo*	*Cyclanthera pedata*
Genome Size (bp)	156,906	155,402	155,293	157,533	157,204	157,592	157,343	155,027
SSC length (bp)	86,845	86,287	89,668	88,113	87,652	88,182	87,971	84,165
LSC length(bp)	17,896	18,086	22,910	18,143	18,039	18,181	18,167	18,289
IRA length (bp)	26,081	25,513	21,356	25,637	25,755	25,613	25,601	26,285
IRB length (bp)	26,081	25,513	21,356	25,637	25,755	25,613	25,601	26,285
No. of different protein-coding genes	81	81	81	81	81	81	81	81
No. of different rRNA genes	8	9	10	8	8	8	9	8
No. of different tRNA genes	38	38	41	38	38	38	38	41
No. of different genes	127	128	132	127	127	127	128	130
No. of different genes with introns	9	9	9	9	9	9	9	9
No. of codons	26,370	25,309	26,052	25,545	26,626	26,625	25,275	26,588
Coding sequence length (bp)	79,110	75,926	78,155	76,634	79,879	79,876	75,826	79,765
No. of genes present in both IR	23	26	20	23	23	23	22	23
%GC content in LSC	34.94	34.67	35.03	34.90	34.85	34.89	34.91	35.28
%GC content in SSC	31.54	30.94	33.47	31.55	31.47	31.44	31.44	30.95
%GC content in IR	42.84	42.79	43.32	43.00	42.95	43.01	43.05	42.80

Comparative analysis of genome structure

The structural characteristics of the cp genome of *C moshata* was explored, and compared it with other seven Cucurbitaceae species: *Citrullus lanatus, Cucumis melo, Cucumis sativus, Cucurbita ficifolia, C. maxima, C. moschata, C. pepo* and *Cyclanthera pedata*. The cp genomes of these species differed in 686 pb, 2,190 pb, 2,299 pb, 59 pb, 388 pb, 249 pb, and 2565 pb, respectively (Table 1). Using mVISTA program, the divergence in the chloroplast genomes of seven cucurbit species was identified. The identity between the sequences and colored regions of high conservation is visualized in Fig. 2. As expected, the IR regions showed less divergence than the SSC and LSC regions. An identity analysis between the cp of *C. moschata* and seven closely related species showed higher percentage of identity is evidenced between *C. moschata* and *C. maxima* (0.99) and higher divergence values with *Cucumis Sativus* (0.68) (Figure S1).

A genome-wide analysis by sliding window assessment to detect hotspot regions in *Cucurbita* cp genomes was conducted. By applying a Pi-value cut-off threshold of 0.055, it was possible to detect six gene regions with a high degree of variability. Of these, three regions were located within the LSC (*trnG-TCC*, *trnT-GGT* and *ndhC*) and three within the SSC (*ndhF*, *rpl32* and *ycf1*) region (Fig. 3). Noncoding regions included: *trnT-GGU–psbD*, *trnR-UCU–atpA, trnL-UAA–trnF-GAA, accD–pasI, and ndhF–rpl32*, *atpH–atpI* (Fig. 2).

Table 2
List of genes found in cp genome of “loche”, *C. moschata*

Category	Function	Genes
RNA genes	Transfer RNA	**trnH-GUG**, **trnK-UUU**[b], **trnQ-UUG**, **trnS-GCU**, **trnG-UCC**[b], **trnR-UCU**, **trnC-GCA**, **trnD-GUC**, **trnY-GUA**, **trnE-UUC**, **trnT-GGU**, **trnS-UGA**, **trnG-GCC**, **trnfM-CAU**, **trnS-GGA**, **trnT-UGU**, **trnL-UAA**[b], **trnF-GAA**, **trnV-UAC**[b], **trnM-CAU**, **trnW-CCA**, **trnP-UGG**, *trnI-CAU (x2), trnL-CAA*[b]*(x2), trnV-GAC (x2), trnI-GAU*[b] *(x2), trnA-UGC*[b] *(x2), trnR-ACG (x2), trnN-GUU (x2)*, **trnL-UAG**
	Ribosomal RNA	*rrn23 (x2), rrn16 (x2), rrn5 (x2), rrn4.5(x2)*
Transcription and translation related genes	Transcription and splicing	**rpoA**, **rpoB**, **rpoC1**[b], **rpoC2**
	Ribosomal protein large subunit	*rpl2*[b] *(x2)*, **rpl14**, **rpl16**[b], **rpl20**, **rpl22**, *rpl23 (x2)*, **rpl32**, **rpl33**, **rpl36**
	Ribosomal protein small subunit	**rps2**, **rps3**, **rps4**, *rps7 (x2)*, **rps8**, *rps12*[ac] *(x2)*, **rps11**, **rps14**, **rps15**, **rps16**[b], **rps18**, **rps19**
Photosynthesis	ATP synthase	**atpA**, **atpB**, **atpE**, **atpF**[b], **atpH**, **atpI**
	Photosystem I	**psaA**, **psaB**, **psaC**, **psaI**, **psaJ**
	Photosystem II	**psbA**, **psbB**, **psbC**, **psbD**, **psbE**, **psbF**, **psbH**, **psbI**, **psbJ**, **psbK**, **psbL**, **psbM**, **psbT**, **psbZ**
	Cytochrome complex	**petA**, **petB**[b], **petD**, **petG**, **petL**, **petN**
	Calvin cycle	**rbcL**
	NADH dehydrogenase	**ndhA**[b], *ndhB*[b] *(x2)*, **ndhC**, **ndhD**, **ndhE**, **ndhF**, **ndhG**, **ndhH**, **ndhI**, **ndhJ**, **ndhK**
Miscellaneous group	Translation initiation factor IF1	**infA**
	Acetyl-CoA carboxylase	*accD*
	Maturase	**matK**
	ATP-dependent protease	**ClpP1**[a]
	Inner membrane protein	**cemA**
	Cytochrome c biogenesis	**ccsA**
	Photosystem assembly/stability factors	**Pbf1, pafI**[b], **pafII**
Other genes	Conserved hypothetical chloroplast ORF	**ycf1**, **pseudogen ycf1**, *ycf2 (x2)*

[a]Gene containing two introns; [b]Gene containing a single intron; [c]Gene divided into two independent transcription units

Simple sequence repeats identification

We identified microsatellites or simple sequence repeats (SSRs) in the cp genome of seven related species of Cucurbitaceae. In *C. moschata*, a total of 59 SSR were identified. Mononucleotide repeats (43) were the most abundant, followed by tetranucleotide (6) and dinucleotides (7) SSR with trinucleotides (2), and hexanucleotide (1) repeat motifs were recognized in less quantity. On the other hand, most of the mononucleotide repeats were A/T motifs, while AT/TA motifs from dinucleotide repeats (Fig. 4).

Similarly, 47, 73, 68, 65, 56, 61 and 40 SSRs were found in *Citrullus lanatus, Cucumis melo, Cucumis sativus, Cucurbita ficifolia, Cucurbita maxima, C. moschata, C. pepo* and *Cyclanthera pedata*, respectively (Table S1). On the other hand, only *Cucumis melo* and *Cucurbita moschata* presented hexanucleotide repeats. In addition, *C. maxima*, C. *moschata, C. pepo* and *Cyclanthera pedata* did not present pentanucleotide repeats. We also identified the distribution of SSRs in LSC, IR and SSC regions. Most of the repeats were located in the LSC region, varying from 14 in *Cyclanthera pedata*, 15 in *Citrullus lanatus*, 17 in *Cucurbita maxima*, 19 in *Cucumis pepo*, 20 in *C. moschata*, 21 in *C. ficifolia*, 23 in *Cucumis melo*, to 24 in *C. Sativus*, followed by the IR regions (9 and 7 in *C. moschata*) and SSC region (Fig. S2).

Taxonomical relationships

There is a high percentage identity between genera *Cucurbita* and *Citrullus*, ranging from 0.98 to 0.99. On the contrary, the identity between genera *Cucurbita* and *Gynostemma* revealed minor percentage of identity, 0.33 to 0.34. However, the species that showed the lowest percentage of identity is *Cyclanthera pedata* with: (i) *Gynostemma pentaphyllum*, (ii) *Hemsleya zhejiangensis*, (iii) *G. pentaphyllum*, (iv) *Gynostemma microspermum*, (v) *Gynostemma compressum* and (vi) *Gynostemma cardiospermum* (Fig. S3).

Phylogenetic inference of loche, C. moschata

The maximum likelihood tree showed all nodes possess 100% bootstrap support (BS), except for three (98%, 76% and 54%). Members of the Cucurbitae section were clustered into one group with 100% BS. Similarly, the seven species of the Benincaseae section were placed within one cluster with very high BS. Loche landrace (*C. moschata*) was sister species with *C. pepo* with 100% BS (Fig. 5).

Discussions

In this study, the cp genome of a Peruvian landrace of *C. moschata*, “loche” was sequenced using Illumina sequencing technology. This analysis showed that the cp genome of “loche” conserves a typical quadripartite structure: LSC, SSC separated by inverted repeats regions IRa and IRb. The cp genomes are highly conserved in such features as genomic size and structure [24]. The organization and structure of the *C. moschata* is similar to other sequenced *Cucurbitaceae* [21]. The size of the cp genome is 157,592 bp very similar to *C. ficifolia* with only 59 bp difference [19]. In most studied species the size of the genome varies significantly in the LSC region, in comparison with the IR regions [19, 21] which are highly conserved. The stability of the IR regions maybe is due to their important role in the recombination process; indeed, recombination is implicated in the replication and repair of organelle genomes. Alteration in the cycle of the recombination in organelles produces genomic instability, often accompanied by adverse consequences for plant fitness [25].

The IR region revealed low differentiation among species analyzed (0.96–1), except between *Cucumis melo* and *C. sativa* (0.78), which also differed in 4,157 bp in that region (Table 1). This high degree of differentiation would be supported by integrating different clades shown in phylogeny analyzes [19]. As expected, in agreement with most angiosperms, IRs and coding regions were more conserved than no coding regions [26]. This was consistent with our results based on sequence identity where the variations in the IRs were smaller than the non-coding regions. The alignment of the seven cp genome revealed variable regions which included *trnH-GUG - psbA*, *rps16 - trnQ*, *trnC - petN*, *trnL - trnF*, *rps15 - ycf1*, *rps12 - trnV*, *trnL - trnA*. These regions may be used as makers for identification of Cucurbitaceae species as well as resolving phylogenetic relationships in the family.

The cp genomes of the genus *Cucurbita* encoded 127 genes (except *C. pepo*), containing 81 protein-coding genes, 38 tRNA genes, eight rRNA genes, and one pseudogene. Similar to previously published work in other cucurbits [27, 28, 29], the cp genomes of all *Cucurbita* species possessed a GC content that ranged 34.67–35.28% in the LSC region, 30.94–33.47 in SSC region and 42.80-43.32% in IR region. *Cucumis sativus* and *Cyclanthera pedata* presented the highest GC content; 35.03% and 35.28%, respectively. The highest percentage of GC in the IR region is possibly due to the presence of rRNA in this region. On the other hand, *C. moschata* has one pseudogene, *ycf1*, in the IRb region. This is considered a pseudogene in the chloroplast of many flowering plants because stop codon is absent, therefore it lost its ability to code a protein [30]. This feature differentiates the distribution of single-copy genes and inverted repeat borders. Zou et al [31] suggested that pseudogenes provided important information about gene history and genome evolution as they were evolutionary molecules of functional components in the genome.

The *infA* gene, located between the *rpl36* and *rps8* genes in the LSC region, acts as a translation initiation factor 1 and has RNA chaperone activity in more than 300 diverse angiosperms [32]. However, it is present as a pseudogene in many other cucurbit cp genomes: *Gynostemma yixingense, G. microspermum, Citrullus colocynthis, Cucumis melo* [27, 28, 33, 34] or is lost like in *Arabidopsis thaliana* and *Alstroemeria aurea* [35, 36]. Many genes were lost during the early evolution of photosynthetic eukaryotes, often in parallel in different algal lineages, and some of these losses were the result of gene transfers to the nuclear genome. It has been shown at least 24 losses of *infA* in angiosperms [32] and jointly with *ndhF* are the only genes that have been lost repeatedly, which probably reflects repeated loss of the entire chloroplast NADH dehydrogenase subunit complex [33]. Therefore, *infA* is classified as an unusually unstable gene in the angiosperm chloroplast, making this gene by far the most mobile chloroplast gene known in plants [32].

Microsatellites are widely used as molecular markers based on their polymorphism leading to sensitive genetic diversity, population structure, phylogenetic relationship inference at the inter- and intrapopulation levels [37, 38]. Moreover, SSRs are also related to different types of genome rearrangement, large inversions and recombinations [1]. Thus, “loche” cp SSRs could contribute to evolutionary and molecular ecological knowledge, which warrants further research. Previous studies showed that mononucleotide-type sequences were more abundant in the LSC region compared to SSC and IR regions [39]. Furthermore, a greater number of palindromic repeats were found among four types of repeats, while previous studies revealed that the forward repeats were the most abundant repeats [28]. Phylogenetics guides the study of plant domestication, as it resolves sister relationships between crops and their wild relatives, thus identifying the ancestors of cultivated plants [40]. Our ML tree is well resolved and is in agreement with previous phylogenetic work using complete chloroplast genomes [19, 27, 33, 40, 41]. However, further molecular studies, including additional collections of “loche” from a wider geographic area, are needed to confirm its origin and domestication process.

The information regarding the genome organization, gene content, and structural variation of loche and other cucurbits will help in the conservation of this pre-Columbian landrace of *C. moschata*, which has been cultivated since Chimú culture (1300–1470 d.c.) [42]. In addition, we expect this work may stimulate other researchers to develop molecular tools for other Peruvian neglected crops aiming to develop modern breeding programs in order to alleviate poverty of the rural areas.

Methods

Plant material, DNA isolation and sequencing

Fresh leaves of *C. moschata* were sampled from a commercial field owned by Manuel A. Mesones Muro, (79° 44' 11" O, 6° 38' 44" S) Lambayeque, Peru. The specimen was deposited in the herbarium of Universidad Nacional Mayor de San Marcos (UNMSM), under the voucher N° 337139. The total genomic DNA was isolated from fresh leaves of “loche” using the CTAB method [43], with modifications for this specie. DNA quantity and quality was evaluated through 1% agarose and Qubit™4

Fluorometer (Invitrogen,Waltham, MA, USA), respectively. Pair-end 150 reads were obtained by the Illumina HiSeq 2500 platform using the NexteraXT DNA Library Preparation Kit (Illumina, San Diego, CA, USA).

Genome Assembly and Annotations

Raw reads were filtered with Trim Galore software [44] with default arguments. The cp genome was assembled from high-quality clean reads using GetOrganelle v1.7.2[45] program with the same parameters as followed by Saldaña et al. 2022 [39], using *C. moschata* (accession number: NC_036506) as reference. SPAdes v3.11.1 [46], bowtie2 v2.4.2 [47], and BLAST + v2.11 [48] were also run with default options.

Geseq online tool [49] was used to annotate the genes and program tRNAscan-SE ver 1.21 [50] was employed to detect tRNA genes with default settings. To validate rRNA genes in chloroplast, RNAmmer [51] software was used with default settings. We included plastid genomes of related species within Cucurbitaceae available at NCBI. The chloroplast genome of “loche” was manually curated and the architecture of “loche” cp genome was visualized with OGDRAW 1.3.1 [52].

Cucurbitaceae genomes comparison

MVISTA program with Shuffle-LAGAN model [53] was used to align and compare the cp genome of “loche” with seven species of the Cucurbitaceae family: (i) *Citrullus lanatus*, (ii) *Cucumis melo*, (iii) *C. sativus*, (iv) *Cucurbita ficifolia*, (v) *C. maxima*, (vi) *C. moschata*, (vii) *C. pepo* and (viii) *Cyclanthera pedada*. The annotated genome of *C. moschata* cp genome (accession number: NC_036506) was used as reference. An identity matrix was generated. First, independent alignments of each of the sequences were conducted using MAFFT v7.475 [54], following the same parameters reported by Saldaña et al. [39]. Manual alignment corrections were performed using MacClade v4.08a software. In order to identify areas of high mutational frequency within the plastomes of the eight species under study, MAFFT [54] was employed to perform sequence alignment of the cp sequences. Subsequently, DnaSP v5 [55] was utilized to calculate nucleotide variability (Pi) among the cp genomes using a window length of 800 bp and a step size of 200 bp. Finally, to determine identity values, an identity plot was created using the ggplot2 [56], gtext (https://github.com/wilkelab/ggtext/) and ggpubr [57] package in the R software [58].

SSR analysis in the chloroplast genome of Cucurbitacea

MIcroSAtellite (MISA) [59] software was employed to identify the simple sequence repeats (SSRs) in the seven Cucurbitaceae cp genomes. The repeats for mononucleotides, dinucleotides, trinucleotides, tetranucleotides, pentanucleotides, and hexanucleotides were 10, 5, 4, 3, 3, and 3, respectively [60]. Also, a plot with the structure and location of the SSRs in seven cp genomes was generated using gggenomes (https://github.com/thackl/gggenomes) and genoPlotR [61] packages in the R software.

Taxonomical relationships

Aiming to investigate the taxonomical relationships within Cucurbitaceae based on the whole plastome sequences, we employed the average nucleotide identity (ANI) analysis. ANI was calculated for the whole cp genomes using the Pyani script (Python module) for average nucleotide identity analyses; (https://github.com/widdowquinn/pyani), aligning the sequences with the MUMmer algorithm [62].

Phylogeny of C. moschata

RAxML v8.2.11 software was used to construct a maximum likelihood (ML) phylogenetic tree with 1000 nonparametric bootstrap replicates under the GTR + nucleotide substitution model of evolution. The 28 cp genomes selected from the Organelle Genome Resources of the NCBI were compared and aligned by MAFFT program with “loche” cp genome. *Solanum muricatum* (accession number: OK326864) was included as an outgroup. The resulting tree was viewed with FigTree version 1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

Declarations

Ethics approval and consent to participate

Plant material was collected following relevant institutional and national guidelines.

Clinical trial number

Not applicable

Consent for publication

Not applicable.

Availability of data and material

Sequence data that support the findings of this study have been deposited in the National Center for Biotechnology Information with the primary accession code OQ442842.

Funding

This research was funded by the project “Creación del servicio de agricultura de precisión en los Departamentos de Lambayeque, Huancavelica, Ucayali y San Martín 4 Departamentos” of the Ministry of Agrarian Development and Irrigation (MIDAGRI) of the Peruvian Government with grant number CUI 2449640.

Competing interests

The authors declare no conflict of interest.

Authors' contributions

Conceptualization, C.L.S. and C.I.A.; Methodology, C.L.S., R.E. and C.I.A. ; Software, C.L.S. and R.E.; Validation, C.L.S., R.E. and E.S.; Formal analysis, C.L.S., R.E., and C.I.A.; Investigation, C.L.S., R.E., E.S., C.C., E.V. and C.I.A.; Resources, C.C. and C.I.A.; Data curation, C.L.S., R.E. and E.S.; Writing—original draft, C.L.S., R.E., E.S., C.C., E.V. and C.I.A.; Writing—review & editing, C.L.S., R.E., E.S., C.C., E.V. and C.I.A.; Supervision, C.I.A.; Project administration, C.I.A.; Funding acquisition, C.I.A. All authors reviewed the manuscript.

Acknowledgements

We thank Ivan Ucharima for image editing. In addition, we thank Eric Rodriguez, Maria Angélica Puyo and Cristina Aybar for supporting the logistic activities in our research center. The authors also thank the Bioinformatics High-performance Computing server of Universidad Nacional Agraria la Molina (BioHPC-UNALM) for providing resources to perform the analyses. C.I.A. thanks Vicerrectorado de Investigación of UNTRM.

References

1. José Blanca, Joaquín Cañizares, Cristina Roig, Pello Ziarsolo, Fernando Nuez BP. Transcriptome characterization and high throughput SSRs and SNPs discovery in *Cucurbita pepo* (Cucurbitaceae). BMC Genomics. 2011;12.
2. Salehi B, Quispe C, Sharifi-Rad J, Giri L, Suyal R, Jugran AK, et al. Antioxidant potential of family Cucurbitaceae with special emphasis on Cucurbita genus: A key to alleviate oxidative stress-mediated disorders. Phyther Res. 2021; 35:3533–57.

3. Schaefer H, Heibl C, Renner SS. Gourds afloat: A dated phylogeny reveals an Asian origin of the gourd family (Cucurbitaceae) and numerous oversea dispersal events. Proc R Soc B Biol Sci. 2012; 276:843–51.
4. Smith BD. Reassessing Coxcatlan Cave and the early history of domesticated plants in Mesoamerica. Proc Natl Acad Sci U S A. 2005; 102:9438–45.
5. Shokrzadeh M, Azadbakht M, Ahangar N, Hashemi A, Saeedi Saravi SS. Cytotoxicity of hydro-alcoholic extracts of Cucurbita pepo and Solanum nigrum on HepG2 and CT26 cancer cell lines. Pharmacogn Mag. 2010; 6:176–9.
6. Mukherjee, P. K., Seha S., Amit K., Joydeb C., Subhadip B., Barun D., Pallab K. H., Nanaocha S. Therapeutic Importance of Cucurbitaceae: A Medicinally Important Family. Journal of Ethnopharmacology. 2022; 282:114599.
7. Rolnik, A., Olas B. Vegetables from the Cucurbitaceae family and their products: Positive effect on human health. Nutrition. 2020; 78: 110788.
8. Fita A, Picó B, Roig C, Nuez F. Performance of Cucumis melo ssp. agrestis as a rootstock for melon. J Hortic Sci Biotechnol. 2007; 82:184–90.
9. Ajuru M, Nmom F. A review on the economic uses of species of Cucurbitaceae and their sustainability in Nigeria. American Journal of plant biology. 2017; 2:17-24.
10. Decker-Walters, D.S., Walters TW. Squash. In: Klipe, K.F., Ornelas, K.C. (Eds.), The Cambridge World History of Food. Cambridge University Press, Cambridge. :335–351.
11. Andres, T.; Ugás, R.; Bustamente, F. Loche: A Unique Pre-Columbian Squash Locally Grown in North Coastal Peru; Holmes, G.J., Ed.; Proc. Cucurbitaceae; Universal Press: Raleigh, NC, USA, 2006; 333–340.
12. Arbizu CI, Blas RH, Ugás R. Genetic diversity and population structure assessed by SSR in a Peruvian germplasm collection of loche squash (Cucurbita moschata, Cucurbitaceae). InBiology and Life Sciences Forum. 2022; 15:1-6.
13. Oldenburg DJ, Bendich AJ. DNA maintenance in plastids and mitochondria of plants. Front Plant Sci. 2015;6 OCTOBER:1–15.
14. Daniell H, Lin CS, Yu M, Chang WJ. Chloroplast genomes: Diversity, evolution, and applications in genetic engineering. Genome Biol. 2016; 17:1–29.
15. Amiryousefi A, Hyvönen J, Poczai P. The chloroplast genome sequence of bittersweet (Solanum dulcamara): Plastid genome structure evolution in Solanaceae. PLoS One. 2018; 13:1–23.
16. Zhao Z, Wang X, Yu Y, Yuan S, Jiang D, Zhang Y, et al. Complete chloroplast genome sequences of Dioscorea: Characterization, genomic resources, and phylogenetic analyses. PeerJ. 2018; 2018:1–17.
17. Powell W, Morgante M, Mcdevitt R, Vendramin GG, Rafalski JA. Polymorphic Simple Sequence Repeat Regions in Chloroplast Genomes : Applications to the Population Genetics of Pines Author (s): W . Powell , M . Morgante , R . McDevitt , G . G . Vendramin and J . A . Rafalski Source : Proceedings of the National Acade. 2016; 92:7759–63.
18. Bock R, Khan MS. Taming plastids for a green future. Trends Biotechnol. 2004; 22:311–8.
19. Cong Z, Cai L, Zhang Y, Su W, Li H, Zhu Q. The complete chloroplast genome sequence of the *Cucurbita ficifolia* Bouché (Cucurbitaceae). Mitochondrial DNA Part B Resour. 2021; 6:3095–7.
20. Straub SC, Parks M, Weitemier K, Fishbein M, Cronn RC, Liston A. Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. American journal of botany. 2012; 99:349-64.
21. Zhang X, Zhou T, Yang J, Sun J, Ju M, Zhao Y, et al. Comparative analyses of chloroplast genomes of cucurbitaceae species: Lights into selective pressures and phylogenetic relationships. Molecules. 2018;23.
22. Hu J, Yao J, Lu J, Liu W, Zhao Z, Li Y, Jiang L, Zha L. The complete chloroplast genome sequences of nine melon varieties (Cucumis melo L.): lights into comparative analysis and phylogenetic relationships. Frontiers in Genetics. 2024; 15:1417266.
23. Cruz W, Cuellar E, Ramos M. Manual De Protocolos. 2019.

24. Sun YX, Moore MJ, Meng AP, Soltis PS, Soltis DE, Li JQ, et al. Complete Plastid Genome Sequencing of Trochodendraceae Reveals a Significant Expansion of the Inverted Repeat and Suggests a Paleogene Divergence between the Two Extant Species. PLoS One. 2013; 8:1–13.
25. Maréchal A, Brisson N. Recombination and the maintenance of plant organelle genome stability. New Phytologist. 2010; 186:299-317
26. Asaf S, Khan AL, Lubna, Khan A, Khan A, Khan G, Lee IJ, Al-Harrasi A. Expanded inverted repeat region with large scale inversion in the first complete plastid genome sequence of Plantago ovata. Scientific reports. 2020; 10:3881.
27. Cheng H, Kong WP, Zhang MM, Hou D. The complete chloroplast genome of Cucumis melo L. 'Shengkaihua' (Cucurbitaceae) and its phylogenetic implication. Mitochondrial DNA Part B Resour. 2020; 5:1253–4.
28. Wang L, Lu G, Liu H, Huang L, Jiang W, Li P, et al. The complete chloroplast genome sequence of gynostemma Yixingense and comparative analysis with congeneric species. Genet Mol Biol. 2020; 43:1–6.
29. Song W, Chen Z, He L, Feng Q, Zhang H, Du G, Shi C, Wang S. Comparative chloroplast genome analysis of wax gourd (Benincasa hispida) with three Benincaseae species, revealing evolutionary dynamic patterns and phylogenetic implications. Genes. 2022; 13:461.
30. Vanin EF. Processed pseudogenes identification. Annu Rev Genet. 1985; 19:253–72.
31. Zou C, Lehti-Shiu MD, Thibaud-Nissen F, Prakash T, Buell CR, Shiu SH. Evolutionary and expression signatures of pseudogenes in Arabidopsis and rice. Plant Physiol. 2009; 151:3–15.
32. Millen RS, Olmstead RG, Adams KL, Palmer JD, Lao NT, Heggie L, Kavanagh TA, Hibberd JM, Gray JC, Morden CW, Calie PJ. Many parallel losses of infA from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. The Plant Cell. 2001; 13:645-58.
33. 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; 8:1–13.
34. Wang W, Shao F, Deng X, Liu Y, Chen S, Li Y, et al. Genome surveying reveals the complete chloroplast genome and nuclear genomic features of the crocin-producing plant Gardenia jasminoides Ellis. Genet Resour Crop Evol. 2021; 68:1165–80.
35. Sato S, Nakamura Y, Kaneko T, ASAMIZU E, Tabata S. Complete Structure of the Chloroplast Genome of thaliana ssc. DNA Res. 1999; 290:283–90.
36. Do HDK, Kim JS, Kim JH. Comparative genomics of four Liliales families inferred from the complete chloroplast genome sequence of Veratrum patulum O. Loes. (Melanthiaceae). Gene. 2013; 530:229–35.
37. Ahmed I, Matthews PJ, Biggs PJ, Naeem M, Mclenachan PA, Lockhart PJ. Identification of chloroplast genome loci suitable for high-resolution phylogeographic studies of Colocasia esculenta (L.) Schott (Araceae) and closely related taxa. Mol Ecol Resour. 2013; 13:929–37.
38. Pauwels M, Vekemans X, Godé C, Frérot H, Castric V, Saumitou-Laprade P. Nuclear and chloroplast DNA phylogeography reveals vicariance among European populations of the model species for the study of metal tolerance, Arabidopsis halleri (Brassicaceae). New Phytol. 2012;193:916–28
39. Saldaña CL, Rodriguez-Grados P, Chávez-Galarza JC, Feijoo S, Guerrero-Abad JC, et al. Unlocking the Complete Chloroplast Genome of a Native Tree Spruceanum , Rubiaceae), and Its Comparative Analysis with Other Ixoroideae Species. 2022.
40. Kates HR, Soltis PS, Soltis DE. Evolutionary and domestication history of Cucurbita (pumpkin and squash) species inferred from 44 nuclear loci. Mol Phylogenet Evol. 2017;111:98–109
41. Cui H, Zhu Z, Lu Z, Ding Z, Zhang C, Luan F. The complete chloroplast genome sequence of the Sechium edule (Jacq.) Swartz. (Cucurbitaceae). Mitochondrial DNA Part B Resour. 2021;6:97–8.
42. INDECOPI 2023. https://www.gob.pe/institucion/indecopi/noticias/845045-loche-de-lambayeque-es-elegido-producto-emblematico-por-los-premios-summum-2023

43. Doyle JJ, Doyle JL. Doyle_plantDNAextractCTAB_1987.pdf. Phytochemical Bulletin. 1987;19:11–5.
44. Martin M(. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2013;17:10–2.
45. Jin JJ, Yu W Bin, Yang JB, Song Y, DePamphilis CW, Yi TS, et al. GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. bioRxiv. 2020;:1–31.
46. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19:455–77.
47. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
48. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: Architecture and applications. BMC Bioinformatics. 2009;10:1–9.
49. Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, et al. GeSeq - Versatile and accurate annotation of organelle genomes. Nucleic Acids Res. 2017;45:W6–11.
50. Lowe TM, Eddy SR. TRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1996;25:955–64.
51. Lagesen K, Hallin P, Rødland EA, Stærfeldt HH, Rognes T, Ussery DW. RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 2007;35:3100–8.
52. Greiner S, Lehwark P, Bock R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: Expanded toolkit for the graphical visualization of organellar genomes. Nucleic Acids Res. 2019;47:W59–64.
53. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: Computational tools for comparative genomics. Nucleic Acids Res. 2004;32 WEB SERVER ISS.:273–9.
54. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.
55. Librado P, Rozas J. DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009;25:1451–2.
56. Wickham H. ggplot2-Elegant Graphics for Data Analysis. Springer International Publishing. Cham, Switz. 2016.
57. Kassambara A. Ggplot2' Based Publication Ready Plots Version. 2017.
58. Team RC. R: A Language and Environment for Statistical Computing; R Foundation for Statistical Computing. Vienna, Austria. 2020;:326864.
59. Beier, S, Thiel T , Münch t, Scholz, U and MM. MISA-web: a web server for microsatellite prediction Sebastian. Bioinformatics. 2017;33:2583–5.
60. Zhang Y, Zhang JW, Yang Y, Li XN. Structural and comparative analysis of the complete chloroplast genome of a mangrove plant: Scyphiphora hydrophyllacea Gaertn. f. and Related Rubiaceae Species. Forests. 2019;10.
61. Guy L, Kultima JR, Andersson SGE, Quackenbush J. GenoPlotR: comparative gene and genome visualization in R. Bioinformatics. 2011;27:2334–5.
62. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol. 2007;57:81–91.

Figures

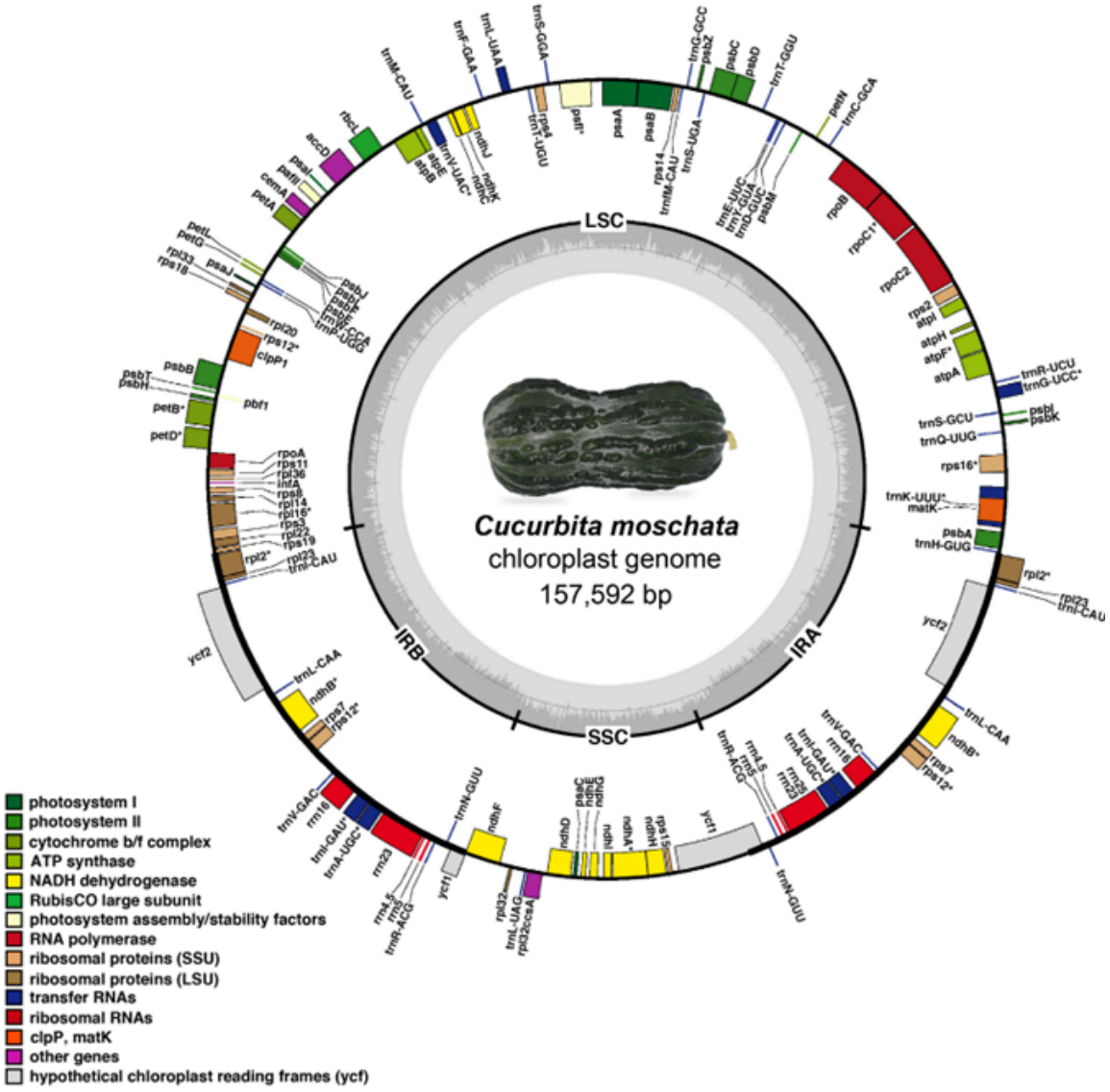

Figure 1

Chloroplast genome map of “loche”, *C. moschata*. Genes outside and inside the circles are transcribed in counterclockwise and clockwise direction. Colored bar indicated functional. Grey and light grey color in the inner circle respectively shows the GC and AT content. LSC indicates large single copy; SSC, indicates small single copy and IR, indicates inverted repeat

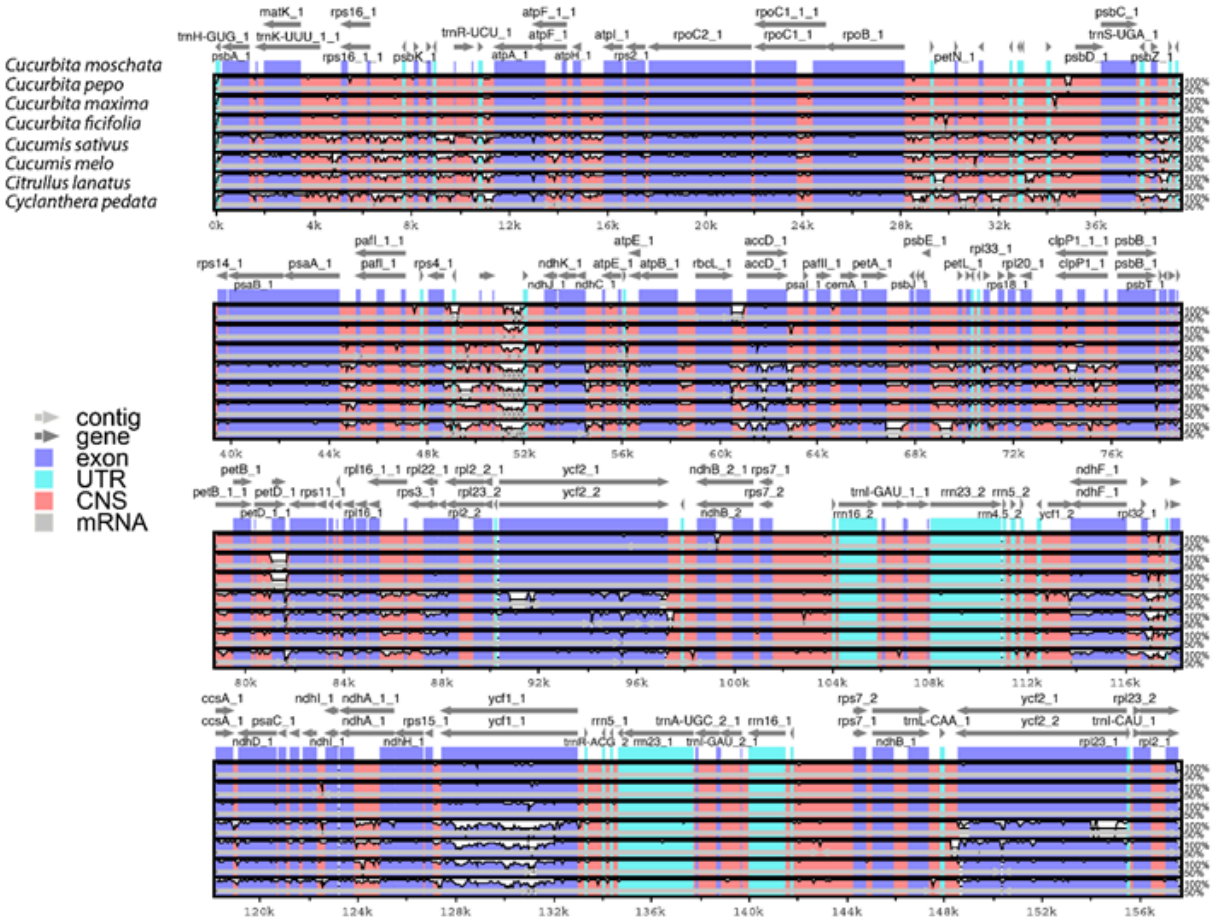

Figure 2

Identity plot comparison of the chloroplast genome of with seven cp genomes. The x-axis represents the coordinate in the chloroplast genome. The percent identity ranging from 50–100% is presented on the y-axis. Genome regions are color-coded.

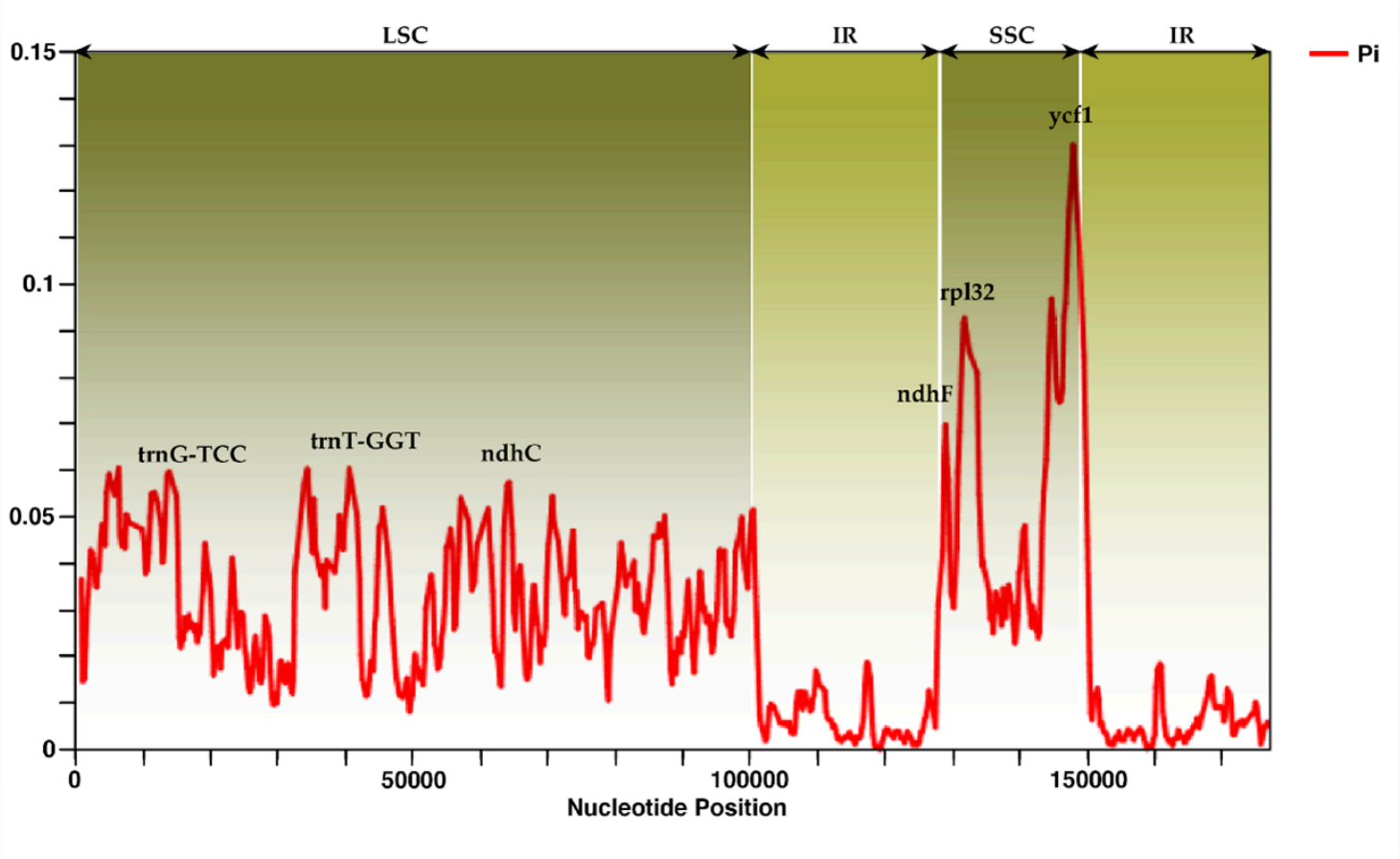

Figure 3

Sliding window analysis of complete cp genome sequences among species of Cucurbitaceae (window length: 800 bp; step size: 200 bp)

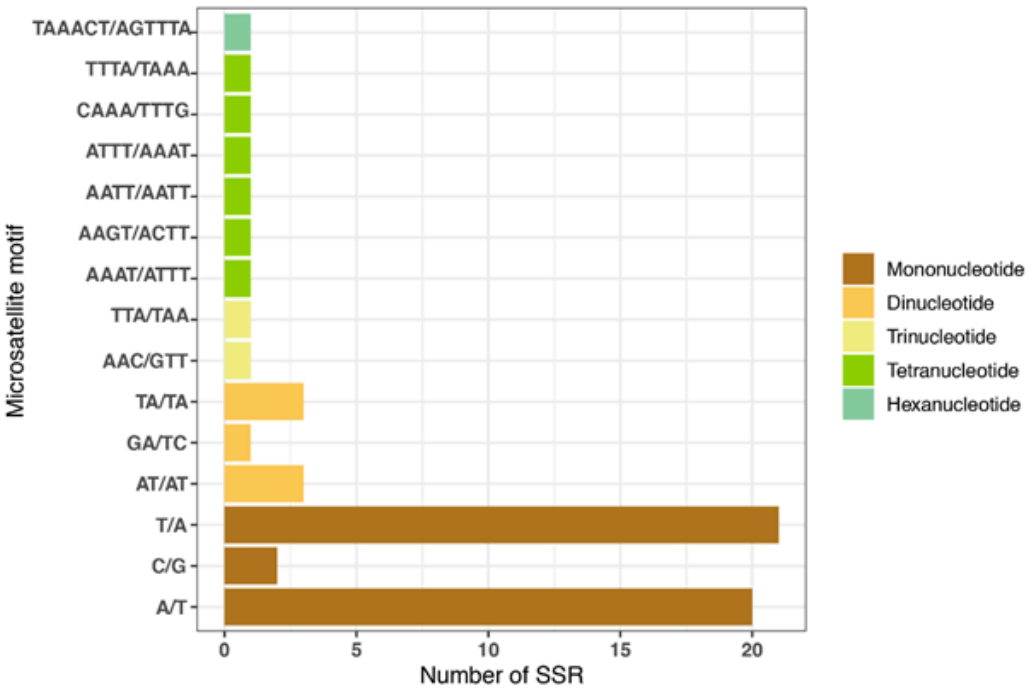

Figure 4

Number and types of microsatellites in *C. moschata*

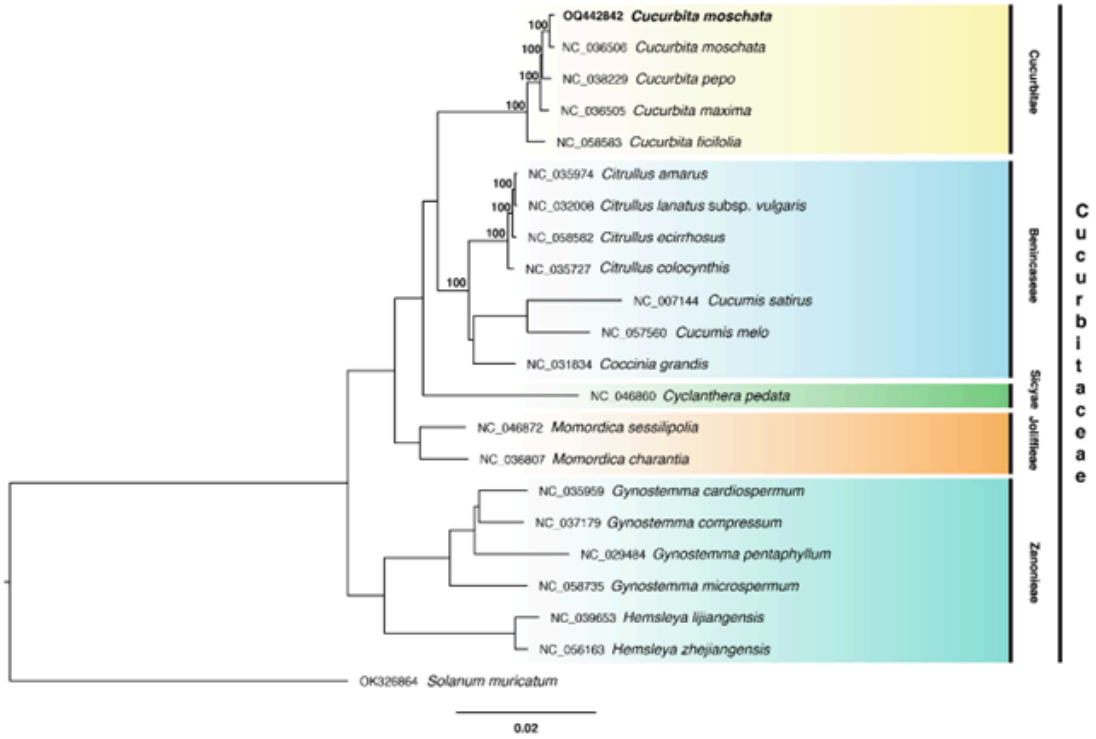

Figure 5

Maximum likelihood tree of the Cucurbitaceae family based on 21 complete chloroplast genome data. The numbers above the nodes represent bootstrap values. The outgroup taxon is *Solanum muricatum*.

Supplementary Files

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

- FigureS1.pdf
- FigureS2.pdf
- FigureS3.pdf
- TableS1.xlsx