

# The complete plastomes of red fleshed pitaya (*Selenicereus monacanthus*) and three related *Selenicereus* species: insights into gene losses, IR expansions and phylogenomic implications

Qiulin Qin

Southwest University of Horticulture and Landscape

Jingling Li

Southwest University of Horticulture and Landscape

Siyuan Zeng

Southwest University of Horticulture and Landscape

Yiceng Xu

Southwest University of Horticulture and Landscape

Fang Han

Southwest University of Horticulture and Landscape

Jie Yu (✉ [yujie1982@swu.edu.cn](mailto:yujie1982@swu.edu.cn))

Southwest University of Horticulture and Landscape

---

## Research Article

**Keywords:** plastome, Hylocereus, Selenicereus, gene, phylogenomics

**Posted Date:** April 1st, 2021

**DOI:** <https://doi.org/10.21203/rs.3.rs-373541/v1>

**License:**   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

---

# Abstract

**Background:** *Selenicereus* is a genus of perennial shrub from the family Cactaceae, and some of them play an important role in the food industry, pharmaceuticals, cosmetics and medicine. To date, there are few reports on *Selenicereus* plastomes, which limits our understanding of this genus. Here, we reported the complete plastomes of four *Selenicereus* species (*S. monacanthus*, *S. anthonianus*, *S. grandiflorus* and *S. validus*), and carried out a comprehensive comparative analysis.

**Results:** The four *Selenicereus* plastomes all have a typical quartile structure. The plastome size ranged from 133,146 bp to 134,450 bp, and contained 104 unique genes, including 30 tRNA genes, 4 rRNA genes and 70 protein-coding genes. Comparative analysis showed that there were massive losses of *ndh* genes in *Selenicereus*. Besides, we observed the IR regions had undergone a dramatic expansion and formed a previously unreported SC/IR border in the intron region of the *atpF* gene. Furthermore, we identified 6 hypervariable regions (*trnF-GAA-rbcL*, *ycf1*, *accD*, *clpP-trnS-GCU*, *clpP-trnT-CGU* and *rpl22-rps19*) that could be used as potential DNA barcodes for the identification of *Selenicereus* species. Phylogenetic analysis indicated that *Hylocereus* was nested in *Selenicereus*.

**Conclusion:** Our study enriches the plastomic resources in the family Cactaceae, and provides the basis for the reconstruction of phylogenetic relationships.

## Background

*Hylocereus* species are perennial herbs from the family Cactaceae. The species in this genus are native to Central America, and nearly 20 species of *Hylocereus* are recognized by most and they can be found naturally occurring from Southern Mexico down throughout Central America and into Northern South America [1]. Also, for harvest their large fruits, which are known as dragon fruit, those species are grown on farms in other parts of the world, especially tropical Asia. All *Hylocereus* species have varying edible fruits, and are commercially developed in different ways. Although the white pitaya (*H. undatus*) is the primary species found in grocery stores and street markets, red-fleshed dragon fruit has gained more popularity. The red fleshed pitaya (*Selenicereus monacanthus* (Lem.) D.R.Hunt), formerly known as *H. lemairei*, not only has an attractive red-purple appearance and unique taste, due to its rich content of high-value functional compounds [2], it is also widely used in pharmaceutical, cosmetic and medical applications. For example, the pulp of red-fleshed pitaya is rich in  $\beta$ -carotene and anthocyanin, which can effectively prevent and treat some chronic diseases (especially cancer) [3–5].

The specific definitions of *Hylocereus* and *Selenicereus* have always been controversial [6]. Britton & Rose [7] gave early definitions of the two based on morphology, but Bauer [8] believed that “the transfer of *Selenicereus* to *Hylocereus*” made Britton & Rose's classification concept no longer applicable. Finally, based on a large number of plastid and nuclear DNA sequences, morphology and anatomical data, it was proved that the two genera were not separated, and *Hylocereus* was nested in *Selenicereus* [9–12]. The increasing quality and widespread cultivation of pitaya, as well as different viewpoints and limited

genomic information, have further complicated the taxonomic definition of this genus. Therefore, it is very important to explore the phylogenetic relationship of the *Selenicereus* species based on genomics. Unfortunately, there are few studies on the phylogenetic relationship between *Hylocereus* and *Selenicereus* based on the complete plastid genomes [13].

Organelle genome sequencing can effectively solve the closely related phylogenetic relationships among species [14]. Chloroplasts are an important organelle in plants, which had its semi-autonomous genetic system, known as chloroplast genome or plastid genome (plastome) [15], most plastomes in angiosperms are a typical quadripartite structure [16], consisting of two inverted repeats (IRa and IRb) and two copy regions (LSC and SSC) [17], and the size of the plastomes ranges from 72 to 220 kb [18], including about 110–130 unique genes, of many are involved in photosynthesis [19]. Plastid genomes have been widely used in taxonomic and evolutionary studies [20] due to their small size, simple structure and maternal inheritance [21, 22]. Entire plastid genomes and nuclear DNA clusters are important in distinguishing between closely related or recessive species [23–25]. Besides, although the plastid genomes are generally conserved in terms of sequence differences and structural organization, some non-coding regions may experience an unexpectedly high frequency of nucleotide substitutions, and these hypervariable regions could be used as DNA barcodes for species identification.

In this study, we sequenced, assembled and analyzed the plastid genomes of four *Selenicereus* species, including the red-fleshed pitaya (*S. monacanthus*, formerly classified as *Hylocereus*) and three traditional *Selenicereus* species (*S. anthonyanus*, *S. grandifloras* and *S. validus*). Our main tasks were as follows: (1) we provide four high-quality references *Selenicereus* plastomes; (2) we analyzed the structural characteristics and sequence divergence of the plastomes in *Selenicereus*; (3) we identified simple sequence repeats (SSRs) loci and repeat sequences for further studies on population genetic structure; (4) we inferred the phylogenetic relationships of *Selenicereus* and *Hylocereus* in Cactaceae based on the complete plastome sequence; and (5) we identified the hypervariable regions that could be used as DNA barcodes for commercial identification of pitaya varieties.

## Results

### Overall organization and features of the four plastomes

The plastomes size of these four taxa ranged from 133,146 bp (*S. monacanthus*) to 134,450 bp (*S. validus*). They were typical quadripartite structure, consisting of a large single-copy region, (LSC, 68,076 – 68,877 bp), a small single-copy region, (SSC, 21,716 – 22,023 bp), and a pair of inverted repeat region (IRs, 21,674 – 21,775 bp). Figure 1 showed the plastid genome map. In addition to the differences in length, the GC content of these conserved plastomes also showed slight changes. According to the analysis, the GC content in the four plastomes ranged from 36.29% to 36.43 %, and the GC content in SSC region (39.39% – 39.69 %) was significantly higher than that in LSC region (36.22% – 36.36 %) and IR region (34.83% – 34.98 %) (Table 1).

Table 1  
Plastome features of the four *Selenicereus* species.

Species		<i>S. monacanthus</i>	<i>S. anthonyanus</i>	<i>S. grandiflorus</i>	<i>S. validus</i>
Accession number		MW553055	MW553068	MW553069	MW553070
Length (bp)	Total	133,146	133,317	134,211	134,450
	LSC	68,076	68,203	68,839	68,877
	SSC	21,716	21,766	22,014	22,023
	IR	21,677	21,674	21,679	21,775
GC content (%)	Total	36.40	36.43	36.34	36.29
	LSC	36.25	36.36	36.24	36.22
	SSC	39.69	39.54	39.40	39.39
	IR	34.98	34.98	34.95	34.83
Gene numbers	Total	104	104	104	104
	tRNA	30	30	30	30
	rRNA	4	4	4	4
	Protein-coding	70	70	70	70

Similar to previous reports in cacti plastomes, the 11 *ndh* genes in the analyzed plastomes were partially lost, including *ndhA*, *ndhC*, *ndhE*, *ndhF*, *ndhG*, *ndhH*, *ndhI*, *ndhJ* and *ndhK*. The second exon of *ndhB* gene was also lost, and only the first exon remained. By contrast, only the *ndhD* gene was intact. On the whole, the four plastomes were all composed of 104 unique genes, including 30 unique tRNA genes, 4 unique rRNA genes and 70 unique protein-coding genes. Moreover, we observed the loss of the first exon of the *clpP* gene, which might be pseudogenes with the same to gene *ndhB* (Table 2).

Table 2  
Gene composition in the plastomes of *Selenicereus*.

Category of Genes	Group of Genes	Name of Genes
Ribosomal RNA	rRNA	<i>rrn16S, rrn23S, rrn5S, rrn4.5S</i>
Transfer RNA	tRNA	30 unique trna genes
Photosynthesis	Subunits of ATP synthase	<i>atpA (x2), atpA, atpB, atpE, atpF* (x2), atpH, atpI</i>
	Subunits of photosystem II	<i>psbA (x2), psbB, psbC, psbD, psbE, psbF, psbI (x2), psbJ, psbK (x2), psbM, psbN, psbT, psbZ</i>
	Subunits of NADH-dehydrogenase	<i>ndhB<sup>ψ</sup>, ndhD</i>
	Subunits of cytochrome b/f complex	<i>petA, petB*, petD*, petG, petL, petN</i>
	Subunits of photosystem I	<i>psaA, psaB, psaC, psal, psaJ</i>
	Subunit of rubisco	<i>rbcL</i>
Self-replication	Large subunit of ribosome	<i>rpl14, rpl16*, rpl2, rpl20, rpl22, rpl32, rpl33, rpl36</i>
	DNA dependent RNA polymerase	<i>rpoA, rpoB, rpoC1, rpoC2</i>
	Small subunit of ribosome	<i>rps11, rps12*, rps14, rps15, rps16, rps16*, rps18, rps19, rps2, rps3, rps4, rps7, rps8</i>
Other genes	Subunit of Acetyl-CoA-carboxylase	<i>accD</i>
	c-type cytochrom synthesis gene	<i>ccsA</i>
	Envelop membrane protein	<i>cemA</i>
	Protease	<i>clpP<sup>ψ</sup> (x2)</i>
	Translational initiation factor	<i>infA</i>
	Maturase	<i>matK (x2)</i>

Note. (x2) indicates that the gene located in the IRs and thus had two complete copies, \* and \*\* indicate that genes containing one/two introns. <sup>ψ</sup> indicates that it is a pseudogene.

Category of Genes	Group of Genes	Name of Genes
Unknown	Conserves open reading frames	<i>ycf1</i> , <i>ycf1</i> <sup>ψ</sup> , <i>ycf3</i> <sup>**</sup> , <i>ycf2</i> (x2), <i>ycf4</i>
Note. (x2) indicates that the gene located in the IRs and thus had two complete copies, * and ** indicate that genes containing one/two introns. <sup>ψ</sup> indicates that it is a pseudogene.		

Furthermore, we also clearly observed the loss of intron in two genes: *rp12* and *rpoC1*. Due to the loss of some genes, exons and introns, the number of intron-containing genes in the plastomes of *Selenicereus* species were significantly reduced compared to most other non-cactus species. Except for the trans-splicing gene, *rps12*, there were only 5 protein-coding genes (*petB*, *petD*, *rp16*, *rps16* and *atpF*) containing one intron, and only one gene (*ycf3*) containing two introns. Moreover, 5 tRNA genes containing one intron (*trnL-UAA*, *trnT-CGU*, *trnK-UUU*, *trnA-UGC* and *trnE-UUC*).

In the four *Selenicereus* plastomes, there were 10 protein-coding genes (*atpF*, *atpA*, *clpP*, *psbI*, *psbK*, *rps16*, *matK*, *psbA*, *ycf2*, *ycf1*) and 8 tRNA genes (*trnR-UCU*, *trnT-CGU*, *trnS-GCU*, *trnQ-UUG*, *trnK-UUU*, *trnH-GUG*, *trnM-CAU* and *trnL-CAA*) were found located in the IR regions, they all have two copies. Among the protein-coding genes, two genes (*ycf1* and *atpF*) are partially located in IR region. All rRNA is located in the SSC region.

### Repeat And Ssrs Analysis

Microsatellites (simple repeat sequences, SSRs) are usually 6 bp tandem sequences in eukaryotic genomes [26]. Their high polymorphism and codominant inheritance make them popular molecular markers [27, 28], which play an important role in the identification of species and the evaluation of evolutionary relationships [29]. Among the four plastomes, *S. monacanthus* had the largest number of SSRs (74), followed by *S. anthonyanus* and *S. validus* (both were 61), and *S. grandiflorus* (55). Most of these SSRs were homopolymers of A/T mononucleotide, and on average, they accounted for 64.34% of the total SSRs. Dinucleotides (18.85%), tetranucleotides (8.61%), and trinucleotides (3.69%) were followed. Pentanucleotide and hexanucleotide repeats were rare in *Selenicereus* plastomes, accounting for 1.23% and 1.64% of all SSRs, respectively (Fig. 2 and Additional file 1: Table S1).

We detected a large number of dispersed repeats in the four plastomes. A total of 807 dispersed repeats were identified, including 618 forward repeats (with length ranged from 30 to 415 bp), 146 palindromic repeats (30 to 415 bp), 39 reverse repeats (30 to 41 bp), and four complementary repeats of 30 bp in length (Additional file 1: Table S2). The detailed distribution of these dispersed repeats in each plastome was shown in Fig. 2. Notably, the number of forward repeats in *S. grandiflorus* and *S. validus* was significantly greater than that in the other two taxa. The dispersed repeats not only serve as potential markers for rearrangement, but were also crucial for inducing mutations [30].

### Genomic Divergence

Sequence identity analysis based on mVISTA [31] was performed among the 4 plastomes, with the reference being the plastome of *S. validus*. We found that the plastome sequences of the four species were quite conservative, in general, IR regions were more conserved than LSC and SSC regions, and the hypervariable regions were mainly found in non-coding sequences. Nevertheless, several coding-regions showed significant differences in the sequences (Fig. 3), such as *accD*, *clpP*, *ycf1* and *ccsA*. The *accD* gene, in particular showed unusual sequence divergence. In addition, there were significant differences among several non-coding regions: *trnF-rbcL*, *trnM-accD* and *trnN-trnR*.

According to the results of DNA sequence polymorphism obtained by DnaSP (v6.0) [32], we identified six hypervariable regions, there were *trnF-GAA-rbcL* ( $P_i = 0.05567$ ), *ycf1* ( $P_i = 0.059$ ), *clpP-trnS-GCU* ( $P_i = 0.03067$ ), *clpP-trnT-CGU* ( $P_i = 0.03167$ ), *rpl22-rps19* ( $P_i = 0.02067$ ), and the highest  $P_i$  value of *accD* gene, including the intergenic region *trnM-accD*, with  $P_i$  value ranging 0.00667 to 0.167 (Fig. 4). The maximum  $P_i$  value for this region was given in parentheses. The results were similar to those based on mVISTA, suggesting that these regions could be used as a potential DNA barcodes.

We analyzed 67 orthologous genes in the protein-coding regions of the four plastomes. In our study, a total of 19 genes (*atpA*, *matK*, *petD*, *petG*, *petN*, *psaC*, *psal*, *psaJ*, *psbA*, *psbE*, *psbF*, *psbH*, *psbl*, *psbK*, *psbM*, *psbN*, *psbT*, *psbZ*, *rps16*) in the four species were completely conserved, and 27 genes had a mutation rate of less than 1.0%. However, we also found that some protein-coding genes had a high level of mutation (Fig. 5 and Additional file 1: Table S3), for example, the mutation rates of 2 genes were more than 2%, and the mutation rates of 3 genes (*rpl36*, *ycf1* and *rpl22*) were more than 3%. The highest mutation rates were observed in three genes: *rpl32* (12.34%), *accD* (10.05%) and *clpP* (7.44%).

### Contraction And Expansion Of Irs

We analyzed the IR/SC boundaries and their adjacent genes in the four plastomes, and compared them to previously published related plastomes. It is not difficult to find that the IR/SC border and the adjacent genes of *Selenicereus* plastomes were very similar in structural characteristics except for small differences in gene position. However, we found that the IR lengths and IR boundaries of the four plastomes newly reported here were varied greatly from those previously reported in cacti and related species. The length of IR regions was observed more than 20,000 bp in *Opuntia quimilo* and all other reported non-cactus species in the order Caryophyllales. However, it was only 8,530 bp in *Rhipsalis baccifera*, and less than 2,000 bp in most cacti genera, such as *Mammillaria*, *Carnegiea*, *Lophocereus* et al. Here, in our four sequenced *Selenicereus* plastomes, the IR lengths were ranged from 21,674 to 21,775 bp, indicating that the cacti had undergone a drastic expansion/contraction event in the IR region.

Furthermore, we also analyzed the IR boundaries of plastomes in the IR region that over 2,000 bp. As shown in Fig. 6, in two non-cactus species, the *rps19* gene span the LSC/IRb border, and an *rps19* pseudogene was duplicated in the IRa. The same, the *ycf1* gene span the SSC/IRa border, and an *ycf1* pseudogene was duplicated in the IRb.

In *O. quimilo*, the two LSC/IR boundaries were *ycf15-trnV* and *trnV-trnH*, and the two SSC/IR boundaries were *ndhG-trnL* and *ndhG-ndhE*, respectively. By contrast, in *R. baccifera*, the two LSC/IR boundaries were *rpl23-trnI* and *trnI-trnH*, and the two SSC/IR boundaries were inside of *ycf1*. Due to the dynamic changes of IRs, the IR boundaries were also changed in the four *Selenicereus* plastomes. Although the two SSC/IR boundaries were similar to *R. baccifera*, the second exon of *atpF* gained access to the IR regions, the first exon of *atpF* was not. Thus, a previously unreported LSC/IR boundaries at the intron region of *atpF* was formed. This result suggested that the IR boundaries in cacti plastomes were extremely unstable compared with other Caryophyllales plastomes.

## Phylogenetic Analysis Based On Conserved Protein-coding Genes

In this study, we constructed phylogenetic trees by using the 56 shared plastidial genes as datasets. The tree reconstruction based on Maximum Likelihood (ML) method and Bayesian Inference (BI) method had a highly consistent topology. The stable topological structure and high bootstrap/posterior probability support values of each node indicated the reliability of phylogenetic tree (Fig. 7).

The phylogenetic analysis involved 15 species of subfamily Cereoideae and two outgroups (*Pereskia sacharosa* and *O. quimilo*). In our tree, the four *Selenicereus* species form a monophyletic clade supported by strong support values. The red-fleshed pitaya (*S. monacanthus*) was most closely related to *S. anthonyanus* compared to other two *Selenicereus* species. This result also suggested that *Hylocereus* and *Selenicereus* were paraphyletic groups, the genus *Hylocereus* belonged to *Selenicereus* based on the 56 conserved plastidial genes.

## Discussion

### Changes in the content of plastomes: gene gain/loss and intron loss

In this study, we reported the complete plastid genomes of four *Selenicereus* species. According to the assembly results, the plastomes of these four taxa were typical quartile structure, with a pair of inverted repeats separated by a large-copy region and a small-copy region [33]. The *ndh* genes in plastids play an important role in Circulating Electron Flow (CEF) in the photosystem of most land plants, CEF is attributed to plant maintenance of effective photosynthesis, water stress and light protection [34, 35]. Exciting, we found two interesting phenomena in this study. Firstly, the phenomenon of massive losses of *ndh* genes in the plastome was observed, which was similar to the report by Sanderson et al. [36], only the *ndhD* gene was relatively complete; Secondly, we found that the number of intron-containing genes in the plastomes of *Selenicereus* species was significantly reduced. The main reason for this phenomenon is the losses of exons (*ndhB* and *clpP*) and introns (*rpl2* and *rpoC1*). Introns can effectively improve the expression level of genes under certain conditions, and play an indispensable role in regulating gene expression [37]. This loss has also been observed in other plastomes of subfamily Cereoideae [36, 38, 39], and it probably is a feature unique to this clade.

SSRs and the repeats are crucial for the plastome rearrangement, and are widely used to detect population genetic diversity [40], as well as being considered as markers for DNA fingerprinting [41]. We analyzed the SSRs and repeat sequences in the four plastomes. First, the number of SSRs ranged from 55 to 74, of many were mononucleotide (A/T) polymer, and accounting for 65.57% of all SSRs. This is one of the reasons for the low GC content in the plastome. Second, compared with SSRs there were a lot of dispersed repeats in the four analyzed plastomes, and the length of forward/palindromic repeats even more than 400 bp. The repeated sequences have previously been reported to have the potential to form secondary structures, they can be used to identify the recombination process [42]. In our study, these large numbers of short dispersed repeats most likely facilitated the plastome rearrangement. Unfortunately, our Illumina short reads have not been able to confirm this, and the long reads will be needed to confirm the presence of genomic recombination in the future.

### **The Expansion Of Irs Resulted In A Rare Boundary**

The contraction and expansion of IRs are common in angiosperms [43], which is also one of the factors affecting the length of plastid genome [44]. According to the comparative analysis results, we found that the length of IR regions in the four *Selenicereus* plastomes were exceeded 20 kb. Although this phenomenon also exists in *O. quimilo*, the IR length of most reported genera in cacti such as *Mammillaria*, *Carnegiea* and *Lophocereus* were usually less than 2 kb. Other studies have found that the IR length of *R. baccifera* was only 8530 bp. Apparently, cacti species have been undergone dramatic expansion/contraction events in IR region. Besides, through the analysis of the IR boundaries, we noticed that the positions of each gene in the IR/SC border of the four *Selenicereus* plastomes were not significantly different. However, due to the expansion of IRs, some genes originally located in the LSC region were access to IR region and formed a new IR boundary in the intron region of gene *atpF* that had not been reported before. In general, the IR boundary in cactus family is extremely instability compared to that of other plastomes in Caryophyllales order [45].

### **Hypervariable Regions Were Identified Based On Plastome Sequences**

According to the results of sequence similarity analysis by mVISTA, the four *Selenicereus* plastomes were highly conserved, and there were few regions of difference. The hypervariable regions in plastomes were mainly identified in non-coding regions, which is consistent with the other plastomes in angiosperm [46, 47]. Although there is little difference in plastomes as a whole, some hypervariable regions deserve our attention. Significant differences were observed in some protein-coding genes, such as *clpP*, *ycf1*, *ccsA* and *accD*, particular in gene *accD*, the mutation rates were even higher than that of the non-coding region. While in contrast, the gene with the greatest difference among the other plastomes usually was observed in gene *ycf1*. The differences in *accD* genes might be due to the presence of a large number of forward repeats in this region, which tend to mediate genome rearrangement. A large number of repeats in this region have been previously observed in passion fruit [48], leading to the rearrangement of plastid genomes. Our results suggested that this region is also highly variable in cactus, and that they probably also contribute to genomic recombination in the genus *Selenicereus*. The gene *accD* and *ycf1* both are indispensable for plant adaptation and leaf development [49, 50], and the high variability of nucleotide

sequences of these two genes might be the result of environmental adaptation during evolution [51–53]. However, whether they cause physiological differences between *Selenicereus* and other cactus plants remains to be seen. On the whole, these "hot spots" of mutations could be used as resources for system biology analysis and identification of DNA barcodes in plants. Our results provide a wealth of genetic information for the identification of species for the development of new DNA barcodes in *Selenicereus* [54].

### **Phylogenomic analysis revealed a close relationship among *Selenicereus* species**

In this study, we have constructed the high resolution phylogenetic tree by using the 56 shared plastidial genes as datasets. The results show undisputed monophyly of the 4 *Selenicereus* species. However, it is worth noting that *S. monacanthus*, once classified as *Hylocereus* (synonym: *H. lemairei*), is more closely related to *S. anthonyanus*, the traditional *Selenicereus* species. Our results support the previous studies, namely the two genera were not separated, and *Hylocereus* was nested in *Selenicereus* [9–12]. However, considering the existence of interspecific or even intergeneric hybridization for *Selenicereus* plants [55], it is one-side to perform phylogenetic inferences about species with hybridization origin based on organelle genomes, as organelles are matrilineal inheritance [56]. The combination of nuclear and organelle genes should be considered and used for phylogenetic inference in the future.

## **Conclusion**

In this study, we reported the complete plastid genomes of four *Selenicereus* species. The plastid genomes of these four species were similar to those of other angiosperms with typical quadripartite structure. In general, the structural changes of in the four plastomes were interesting: 1) Large losses of *ndh* genes; 2) The losses of introns and exons leads to a significant decrease in the number of intron genes; 3) The IR region underwent a dramatic expansion and formed a previously unreported SC/IR border in the intron region of the *atpF* gene. Besides, we identified 6 hypervariable regions that could be used as potential DNA barcodes for the identification of *Selenicereus* species. Our study enriches the plastomic resources in the family Cactaceae, and provides the basis for the reconstruction of phylogenetic relationships.

## **Methods**

### **Sampling, DNA extraction and sequencing**

Fresh stems of the red-fleshed pitaya (*S. monacanthus*) were collected from Yulin, Guangxi, China (22°94'N, 110°49'E). The fresh stems of the other three analyzed *Selenicereus* species were collected from the local flower market of Beibei, Chongqing, China (29°81'N, 106°40'E). They were identified by Professor Jie Yu. These species were cultivated for edible use or ornamental plants, and no permission is required to collect these samples. Our experimental research, including the collection of plant materials, complies with institutional, national or international guidelines. All the samples were deposited in the Herbarium of

Southwest University, Chongqing, China (voucher code: YJ-swu002, YJ-swu027 ~ YJ-swu029). Total genomic DNA was extracted by using the CTAB method [57]. The DNA library with an insert size of 350 bp was constructed using a NEBNext® library construction kit and sequenced by using the HiSeq Xten PE150 sequencing platform. Sequencing produced a total of 6.04–6.85 Gb of raw data per species. Clean data were obtained by using Trimmomatic [58]: we removing the low-quality sequences with more than 5% bases being “N”, and a quality value of  $Q < 19$  accounted for more than 50% of the total base. The detailed sequencing data were showed in Additional file 1: Table S4.

### **Genome Assembling And Annotation**

The de novo genome assembly from the clean data was accomplished utilizing GetOrganelle (v1.7.3) with a default setting, and NOVOPlasty (v3.8.1) [59] was continue used for those without assembled circular genome. The correctness of the assembly was confirmed by using Bowtie2 (v2. 0.1) [60] to manually edit and map all the raw reads to the assembled genome sequence under the default settings. Detailed assembly information was shown in Additional file 1: Table S5. The plastomes were initially annotated by using GeSeq [61] with two reference genomes (*Carnegieia gigantea*, GenBank: NC\_027618.1 and *Lophocereus schottii*, GenBank: NC\_041727.1). Subsequently, the annotations with problems were manually edited by using Apollo [62], and genome maps were drawn by OGDRAW [63].

### **Repeats And Ssr Analysis**

The GC content was determined by using the cusp program provided by EMBOSS (v6.3.1) [64]. Simple Repeat Sequences (SSRs) were available through the online site MISA (<https://webblast.ipk-gatersleben.de/misa/>). Additionally, REPuter (<https://bibiserv.cebitec.uni-bielefeld.de/reputer/>) was used to calculate palindromic repeats, forward repeats, reverse repeats, and complementary repeats with the following settings: hamming distance of three and minimal repeat size of 30 bp [65].

### **Sequence Divergence Analysis**

The sequence similarity analysis results of the four plastomes we analyzed were obtained in shuffle-LAGAN mode by using the online site mVISTA ([http://genome.lbl.gov/cgi-bin/VistaInput?num\\_seqs=4](http://genome.lbl.gov/cgi-bin/VistaInput?num_seqs=4)). With the help of PhyloSuite (v1.2.1), we extracted the orthologous genes of the four taxa and aligned the sequences by using the plugin MAFFT (v7.313) embedded in PhyloSuite. The percentage of variable sites was calculated based on the comparison of protein-coding genes by MEGA (v6.0) [66]. A sliding window with both window length and step size of 500 bp was set using DnaSP (v6.0) software to obtain nucleotide polymorphisms (Pi) of four plastomes. IRscope was used for visualizing the IR/SC boundaries (<https://irscope.shinyapps.io/irapp/>) and the adjacent genes.

### **Phylogenetic Analysis**

We downloaded 13 plastomes from NCBI, and the phylogenetic tree of 17 species in the cactus family was reconstructed by combining the four plastomes newly reported here. The data sources for phylogenetic analysis were shown in Additional file 1: Table S6. A total of 56 orthologous genes among

the analyzed plastomes were identified and extracted by using PhyloSuite (v1.2.1) [67]. The 56 shared plastid protein-coding genes includes *atpA*, *atpB*, *atpE*, *atpF*, *atpH*, *atpI*, *ccsA*, *cemA*, *clpP*, *infA*, *matK*, *petA*, *petB*, *petD*, *petG*, *petL*, *petN*, *psaA*, *psaB*, *psaC*, *psaI*, *psaJ*, *psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbJ*, *psbK*, *psbM*, *psbN*, *psbT*, *rbcL*, *rpl14*, *rpl16*, *rpl20*, *rpl22*, *rpl2*, *rpoA*, *rpoB*, *rpoC1*, *rpoC2*, *rps11*, *rps12*, *rps14*, *rps15*, *rps19*, *rps2*, *rps3*, *rps4*, *rps7*, *rps8* and *ycf3*. The corresponding nucleotide sequences were aligned by using MAFFT (v7.450) [68] implemented in PhyloSuite. These aligned nucleotide sequences were concatenated, and used to construct the phylogenetic trees by using the maximum likelihood (ML) method implemented in RAxML (v8.2.4). The parameters were “raxmlHPC-PTHREADS-SSE3 -f a -N 1000 -m GTRGAMMA -x 551314260 -p 551314260”. The bootstrap analysis was performed with 1,000 replicates. Bayesian inferences (BI) analysis was performed in MrBayes (v3.2.6) using the Markov Chain Monte Carlo method with 200,000 generations and sampling trees every 100 generations. The first 20% of trees were discarded as burn-in with the remaining trees being used for generating a consensus tree.

## Abbreviations

SSR: Simple sequence repeat; IRs: Inverted repeats; LSC: Large single-copy; SSC: Small single-copy; ML: Maximum-likelihood; BI: Bayesian inference; DnaSP: DNA Sequences Polymorphism; CTAB: Cetyl trimethylammonium bromide; NCBI: National Center for Biotechnology Information; Pi: Nucleotide diversity/polymorphism.

## Declarations

### Acknowledgements

The authors are grateful to the technical support provided by Novogene.

### Author contributions

JY conceived the study and designed experiments; FH collected the samples and extracted DNA for sequencing by using the Illumina platform; YCX assembled and annotated the plastid genomes; SYZ and JLL carried out the comparative chloroplast analysis; QLQ drafted the manuscript. All authors have read and approved the final manuscript.

### Funding

This work was supported by the National Natural Science Foundation of China [31772260] and Chongqing Study Abroad Innovation Project [cx2019052]. The funders were not involved in the study design, data collection, and analysis, decision to publish, or manuscript preparation.

### Author details

<sup>1</sup> College of Horticulture and Landscape Architecture, Southwest University, Chongqing, 400716, China.

<sup>2</sup> Key Laboratory of Horticulture Science for Southern Mountainous Regions, Ministry of Education, Chongqing, 400716, China.

#### Availability of data and materials

The raw sequencing data generated in this study and the four plastid genome sequences were deposited in NCBI (<https://www.ncbi.nlm.nih.gov/>) with accession number: SAMN18357737, SAMN18357760, SAMN18357760, SAMN18357760, MW553055, MW553068, MW553069 and MW553070. All the samples are saved at the Herbarium of Southwest University, Chongqing, China. All other data and material generated in this manuscript are available from the corresponding author upon reasonable request.

#### Ethics approval and consent to participate

The four collected *Selenicereus* species are widely cultivated in China as ornamental or edible fruits. Experimental researches do not include the genetic transformation, preserving the genetic background of the species used, and any other processes requiring ethics approval.

#### Consent for publication

Not applicable.

#### Conflict of interest

The authors declare that they have no conflict of interest.

#### Author details

<sup>1</sup> College of Horticulture and Landscape Architecture, Southwest University, Chongqing, 400716, China.

<sup>2</sup> Key Laboratory of Horticulture Science for Southern Mountainous Regions, Ministry of Education, Chongqing, 400716, China.

## References

1. Nunes E, Sousa A, Lucena C, Silva S, Lucena R, Alves CA, Alves R: Pitaia (*Hylocereus* sp.): Uma revisão para o Brasil. *Gaia Scientia* 2014, 8:90-98.
2. Zhuang Y, Zhang Y, Sun L: Characteristics of fibre-rich powder and antioxidant activity of pitaya (*Hylocereus undatus*) peels. *International Journal of Food Science & Technology* 2012, 47. doi:10.1111/j.1365-2621.2012.02971.x
3. Bai X, Zhang H: P41 Microwave-assisted extraction and HPLC analysis of polyphenols from pitaya peel and its inhibitory effect on human lung cancer cell line A549. *Biochemical Pharmacology* 2017, 139:139-140. doi:10.1016/j.bcp.2017.06.042

4. Guimaraes DAB, De Castro D, de Oliveira FL, Nogueira EM, da Silva MAM, Teodoro AJ: Pitaya Extracts Induce Growth Inhibition and Proapoptotic Effects on Human Cell Lines of Breast Cancer via Downregulation of Estrogen Receptor Gene Expression. *Oxid Med Cell Longev* 2017, 2017:7865073. doi:10.1155/2017/7865073
5. Villalobos M, Schweiggert R, Carle R, Esquivel P: Chemical characterization of Central American pitaya ( *Hylocereus* sp.) seeds and seed oil. *CyTA - Journal of Food* 2012, 10:78-83. doi:10.1080/19476337.2011.580063
6. Cáliz de Dios H: A new subspecies of *Hylocereus undatus* (Cactaceae) from Southeastern México. *Haseltonia* 2009, 11:11-17. doi:10.2985/1070-0048(2005)11[11:ANSOHU]2.0.CO;2
7. Britton NL, Rose JN: The Cactaceae: descriptions and illustrations of plants of the cactus family, vol. 2: Carnegie Institue, Washington 1920, 248 pp.
8. Bauer R: A synopsis of the tribe Hylocereeae F. Buxb. Cactaceae Systematics Initiatives 2003, 17:3-63
9. Arias S, Terrazas T, Arreola-Nava HJ, Vázquez-Sánchez M, Cameron KM: Phylogenetic relationships in *Peniocereus* (Cactaceae) inferred from plastid DNA sequence data. *Journal of plant research* 2005, 118(5):317-328. doi:10.1007/s10265-005-0225-3
10. Gómez-Hinostrosa C, Hernández HM, Terrazas T, Correa-Cano ME: Studies on Mexican Cactaceae. V. Taxonomic notes on *Selenicereus tricae*. *Brittonia* 2014, 66. doi:10.1007/s12228-013-9308-y
11. Plume O, Straub SCK, Tel-Zur N, Cisneros A, Schneider B, Doyle J: Testing a Hypothesis of Intergeneric Allopolyploidy in Vine Cacti (Cactaceae: Hylocereeae). *Systematic Botany* 2013, 38:737. doi:10.1600/036364413X670421
12. Cruz MÁ, Salvador A, Teresa T: Molecular phylogeny and taxonomy of the genus *Disocactus* ( Cactaceae ), based on the DNA sequences of six chloroplast markers. *Willdenowia* 2016, 46(1):145-164. doi:10.3372/wi.46.46112
13. Korotkova N, Borsch T, Arias S: A phylogenetic framework for the Hylocereeae (Cactaceae) and implications for the circumscription of the genera. *Phytotaxa* 2017, 327:1-46. doi:10.11646/phytotaxa.327.1.1
14. Ivanova Z, Sablok G, Daskalova E, Zahmanova G, Apostolova E, Yahubyan G, Baev V: Chloroplast Genome Analysis of Resurrection Tertiary Relict *Haberlea rhodopensis* Highlights Genes Important for Desiccation Stress Response. 2017, 8(204). doi:10.3389/fpls.2017.00204
15. Palmer JD, Jorgensen RA, Thompson WF: CHLOROPLAST DNA VARIATION AND EVOLUTION IN PISUM - PATTERNS OF CHANGE AND PHYLOGENETIC ANALYSIS. *Genetics* 1985, 109(1):195-213.
16. Palmer JD: Comparative organization of chloroplast genomes. *Annual review of genetics* 1985, 19:325-354. doi:10.1146/annurev.ge.19.120185.001545
17. Yang Y, Zhou T, Duan D, Yang J, Feng L, Zhao G: Comparative Analysis of the Complete Chloroplast Genomes of Five *Quercus* Species. *Frontiers in plant science* 2016, 7:959. doi:10.3389/fpls.2016.00959
18. Pervaiz T, Sun X, Zhang Y, Tao R, Zhang J, Fang J: Association between Chloroplast and Mitochondrial DNA sequences in Chinese *Prunus* genotypes (*Prunus persica*, *Prunus domestica*, and

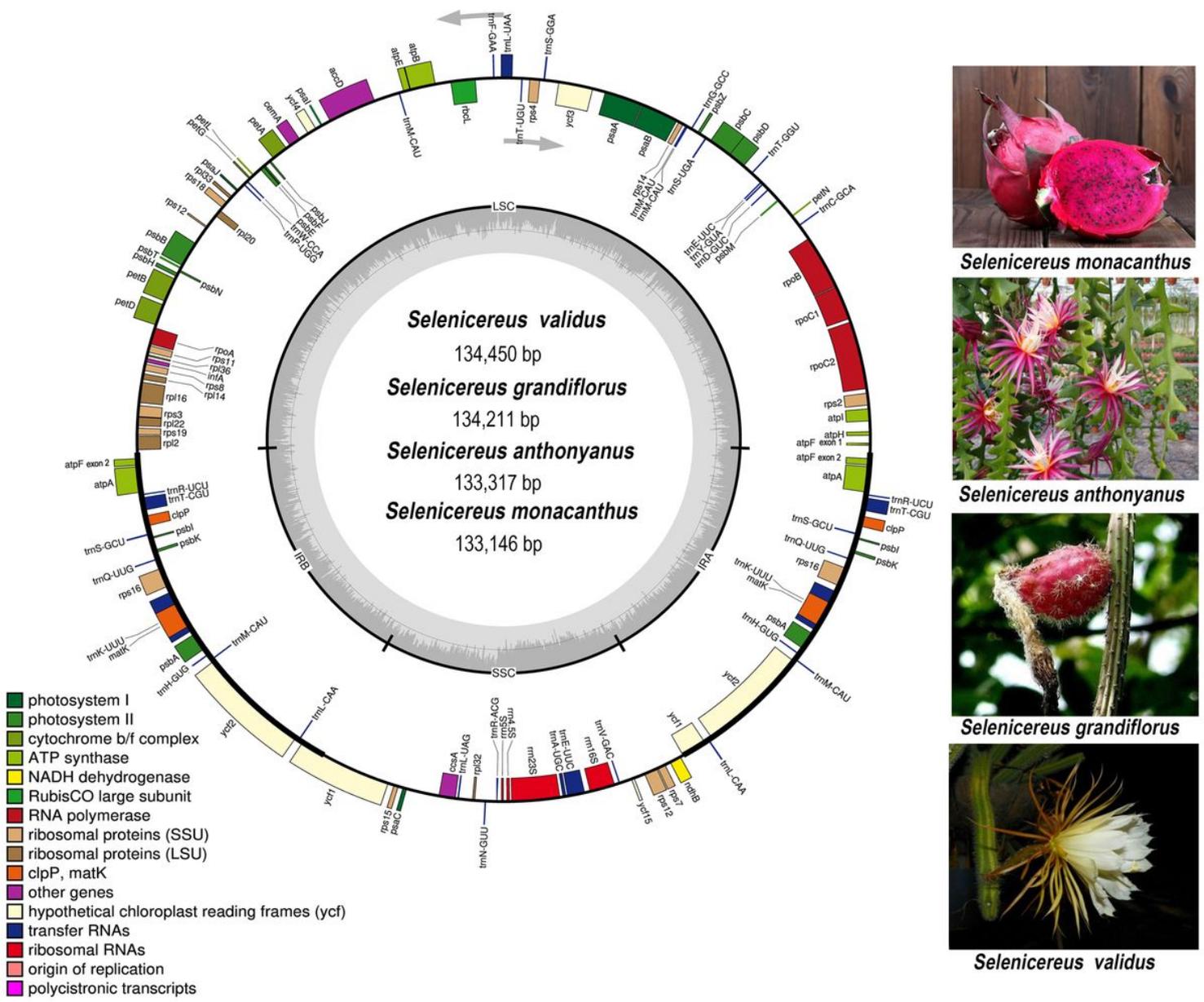
- Prunus avium*). *BMC plant biology* 2015, 15:4. doi:10.1186/s12870-014-0402-4
19. Choi KS, Son OG, Park S: The Chloroplast Genome of *Elaeagnus macrophylla* and trnH Duplication Event in Elaeagnaceae. *PLoS One* 2015, 10(9):e0138727. doi:10.1371/journal.pone.0138727
  20. Daniell H, Lin C-S, Yu M, Chang W-J: Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biology* 2016, 17(1):134. doi:10.1186/s13059-016-1004-2
  21. Maliga P: Engineering the plastid genome of higher plants. *Current opinion in plant biology* 2002, 5(2):164-172. doi:10.1016/s1369-5266(02)00248-0
  22. Palmer JD, Jansen RK, Michaels HJ, Chase MW, Manhart JR: Chloroplast DNA Variation and Plant Phylogeny. *Annals of the Missouri Botanical Garden* 1988, 75(4):1180-1206. doi:10.2307/2399279
  23. Krawczyk K, Nobis M, Myszczyński K, Klichowska E, Sawicki J: Plastid super-barcodes as a tool for species discrimination in feather grasses (Poaceae: Stipa). *Scientific reports* 2018, 8(1):1924. doi:10.1038/s41598-018-20399-w
  24. Yang JB, Yang SX, Li HT, Yang J, Li DZ: Comparative chloroplast genomes of camellia species. *PLoS One* 2013, 8(8):e73053. doi:10.1371/journal.pone.0073053
  25. Myszczyński K, Bączkiewicz A, Buczkowska K, Ślipiko M, Szczecińska M, Sawicki J: The extraordinary variation of the organellar genomes of the *Aneura pinguis* revealed advanced cryptic speciation of the early land plants. *Scientific reports* 2017, 7(1):9804. doi:10.1038/s41598-017-10434-7
  26. Lovin DD, Washington KO, deBruyn B, Hemme RR, Mori A, Epstein SR, Harker BW, Streit TG, Severson DW: Genome-based polymorphic microsatellite development and validation in the mosquito *Aedes aegypti* and application to population genetics in Haiti. *BMC genomics* 2009, 10:590. doi:10.1186/1471-2164-10-590
  27. Morgante M, Hanafey M, Powell W: Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nature genetics* 2002, 30(2):194-200. doi:10.1038/ng822
  28. Naranpanawa DNU, Chandrasekara C, Bandaranayake PCG, Bandaranayake AU: Raw transcriptomics data to gene specific SSRs: a validated free bioinformatics workflow for biologists. *Scientific reports* 2020, 10(1):18236. doi:10.1038/s41598-020-75270-8
  29. Guang XM, Xia JQ, Lin JQ, Yu J, Wan QH, Fang SG: IDSSR: An Efficient Pipeline for Identifying Polymorphic Microsatellites from a Single Genome Sequence. *International journal of molecular sciences* 2019, 20(14). doi:10.3390/ijms20143497
  30. Lopez L, Barreiro R, Fischer M, Koch MA: Mining microsatellite markers from public expressed sequence tags databases for the study of threatened plants. *BMC genomics* 2015, 16:781. doi:10.1186/s12864-015-2031-1
  31. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I: VISTA: computational tools for comparative genomics. *Nucleic acids research* 2004, 32(Web Server issue):W273-279. doi:10.1093/nar/gkh458
  32. Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, Sánchez-Gracia A: DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. *Molecular biology and evolution* 2017, 34(12):3299-3302. doi:10.1093/molbev/msx248

33. Qian J, Song J, Gao H, Zhu Y, Xu J, Pang X, Yao H, Sun C, Li X, Li C *et al*: The complete chloroplast genome sequence of the medicinal plant *Salvia miltiorrhiza*. *PLoS One* 2013, 8(2):e57607. doi:10.1371/journal.pone.0057607
34. Burrows PA, Sazanov LA, Svab Z, Maliga P, Nixon PJ: Identification of a functional respiratory complex in chloroplasts through analysis of tobacco mutants containing disrupted plastid *ndh* genes. *The EMBO journal* 1998, 17(4):868-876. doi:10.1093/emboj/17.4.868
35. Wang P, Duan W, Takabayashi A, Endo T, Shikanai T, Ye JY, Mi H: Chloroplastic NAD(P)H dehydrogenase in tobacco leaves functions in alleviation of oxidative damage caused by temperature stress. *Plant physiology* 2006, 141(2):465-474. doi:10.1104/pp.105.070490
36. Sanderson MJ, Copetti D, Búrquez A, Bustamante E, Charboneau JL, Eguiarte LE, Kumar S, Lee HO, Lee J, McMahon M *et al*: Exceptional reduction of the plastid genome of saguaro cactus (*Carnegiea gigantea*): Loss of the *ndh* gene suite and inverted repeat. *American journal of botany* 2015, 102(7):1115-1127. doi:10.3732/ajb.1500184
37. Yi DK, Lee HL, Sun BY, Chung MY, Kim KJ: The complete chloroplast DNA sequence of *Eleutherococcus senticosus* (Araliaceae); comparative evolutionary analyses with other three asterids. *Molecules and cells* 2012, 33(5):497-508. doi:10.1007/s10059-012-2281-6
38. Oulo MA, Yang JX, Dong X, Wang VO, Mkala EM, Munyao JN, Onjolo VO, Rono PC, Hu GW, Wang QF: Complete Chloroplast Genome of *Rhipsalis baccifera*, the only Cactus with Natural Distribution in the Old World: Genome Rearrangement, Intron Gain and Loss, and Implications for Phylogenetic Studies. *Plants (Basel, Switzerland)* 2020, 9(8). doi:10.3390/plants9080979
39. Solórzano S, Chincoya DA, Sanchez-Flores A, Estrada K, Díaz-Velásquez CE, González-Rodríguez A, Vaca-Paniagua F, Dávila P, Arias S: De Novo Assembly Discovered Novel Structures in Genome of Plastids and Revealed Divergent Inverted Repeats in *Mammillaria* (Cactaceae, Caryophyllales). *Plants (Basel, Switzerland)* 2019, 8(10). doi:10.3390/plants8100392
40. Khan A, Asaf S, Khan AL, Al-Harrasi A, Al-Sudairy O, AbdulKareem NM, Khan A, Shehzad T, Alsaady N, Al-Lawati A *et al*: First complete chloroplast genomics and comparative phylogenetic analysis of *Commiphora gileadensis* and *C. foliacea*: Myrrh producing trees. *PLoS One* 2019, 14(1):e0208511. doi:10.1371/journal.pone.0208511
41. Bodin SS, Kim JS, Kim J-H: Complete Chloroplast Genome of *Chionographis japonica* (Willd.) Maxim. (Melanthiaceae): Comparative Genomics and Evaluation of Universal Primers for Liliales. *Plant Molecular Biology Reporter* 2013, 31(6):1407-1421. doi:10.1007/s11105-013-0616-x
42. Kawata M, Harada T, Shimamoto Y, Oono K, Takaiwa F: Short inverted repeats function as hotspots of intermolecular recombination giving rise to oligomers of deleted plastid DNAs (ptDNAs). *Current genetics* 1997, 31(2):179-184. doi:10.1007/s002940050193
43. Zhu A, Guo W, Gupta S, Fan W, Mower JP: Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. *The New phytologist* 2016, 209(4):1747-1756. doi:10.1111/nph.13743

44. Xue S, Shi T, Luo W, Ni X, Iqbal S, Ni Z, Huang X, Yao D, Shen Z, Gao Z: Comparative analysis of the complete chloroplast genome among *Prunus mume*, *P. armeniaca*, and *P. salicina*. *Horticulture research* 2019, 6:89. doi:10.1038/s41438-019-0171-1
45. Yao G, Jin JJ, Li HT, Yang JB, Mandala VS, Croley M, Mostow R, Douglas NA, Chase MW, Christenhusz MJM *et al*: Plastid phylogenomic insights into the evolution of Caryophyllales. *Molecular phylogenetics and evolution* 2019, 134:74-86. doi:10.1016/j.ympev.2018.12.023
46. Gao X, Zhang X, Meng H, Li J, Zhang D, Liu C: Comparative chloroplast genomes of *Paris Sect. Marmorata*: insights into repeat regions and evolutionary implications. *BMC genomics* 2018, 19(Suppl 10):878. doi:10.1186/s12864-018-5281-x
47. Zhang Y, Du L, Liu A, Chen J, Wu L, Hu W, Zhang W, Kim K, Lee SC, Yang TJ *et al*: The Complete Chloroplast Genome Sequences of Five *Epimedium* Species: Lights into Phylogenetic and Taxonomic Analyses. *Frontiers in plant science* 2016, 7:306. doi:10.3389/fpls.2016.00306
48. Cauz-Santos LA, da Costa ZP, Callot C, Cauet S, Zucchi MI, Bergès H, van den Berg C, Vieira MLC: A Repertory of Rearrangements and the Loss of an Inverted Repeat Region in *Passiflora* Chloroplast Genomes. *Genome biology and evolution* 2020, 12(10):1841-1857. doi:10.1093/gbe/evaa155
49. Kode V, Mudd EA, lamtham S, Day A: The tobacco plastid *accD* gene is essential and is required for leaf development. *The Plant journal : for cell and molecular biology* 2005, 44(2):237-244. doi:10.1111/j.1365-313X.2005.02533.x
50. de Vries J, Sousa FL, Bölter B, Soll J, Gould SB: YCF1: A Green TIC? *The Plant cell* 2015, 27(7):1827-1833. doi:10.1105/tpc.114.135541
51. Park S, Ruhlman TA, Weng ML, Hajrah NH, Sabir JSM, Jansen RK: Contrasting Patterns of Nucleotide Substitution Rates Provide Insight into Dynamic Evolution of Plastid and Mitochondrial Genomes of *Geranium*. *Genome biology and evolution* 2017, 9(6):1766-1780. doi:10.1093/gbe/evx124
52. Thode VA, Lohmann LG: Comparative Chloroplast Genomics at Low Taxonomic Levels: A Case Study Using *Amphilophium* (Bignoniaceae). *Frontiers in plant science* 2019, 10:796. doi:10.3389/fpls.2019.00796
53. de Vries J, Archibald JM, Gould SB: The Carboxy Terminus of YCF1 Contains a Motif Conserved throughout >500 Myr of Streptophyte Evolution. *Genome biology and evolution* 2017, 9(2):473-479. doi:10.1093/gbe/evx013
54. Dong W, Liu J, Yu J, Wang L, Zhou S: Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. *PLoS One* 2012, 7(4):e35071. doi:10.1371/journal.pone.0035071
55. Tel-Zur N, Abbo S, Bar-Zvi D, Mizrahi Y: Genetic relationships among *Hylocereus* and *Selenicereus* vine cacti (Cactaceae): evidence from hybridization and cytological studies. *Ann Bot* 2004, 94(4):527-534. doi:10.1093/aob/mch183
56. Liu X, Wang Z, Shao W, Ye Z, Zhang J: Phylogenetic and Taxonomic Status Analyses of the *Abaso* Section from Multiple Nuclear Genes and Plastid Fragments Reveal New Insights into the North

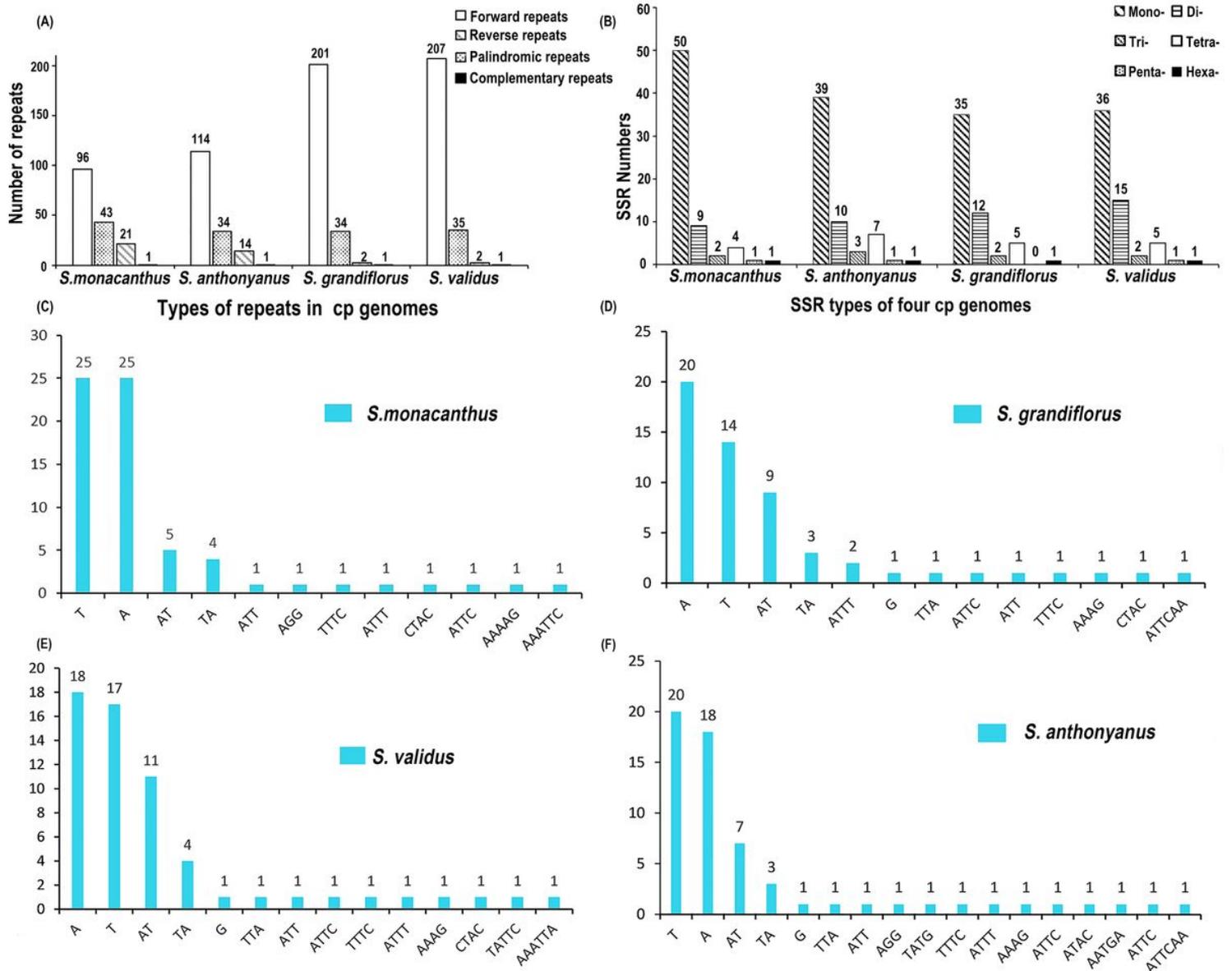
- America Origin of Populus (Salicaceae). *Frontiers in plant science* 2016, 7:2022.  
doi:10.3389/fpls.2016.02022
57. Arseneau JR, Steeves R, Laflamme M: Modified low-salt CTAB extraction of high-quality DNA from contaminant-rich tissues. *Mol Ecol Resour* 2017, 17(4):686-693. doi:10.1111/1755-0998.12616
  58. Bolger AM, Lohse M, Usadel B: Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)* 2014, 30(15):2114-2120. doi:10.1093/bioinformatics/btu170
  59. Dierckxsens N, Mardulyn P, Smits G: NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Research* 2017, 45(4):e18-e18. doi:10.1093/nar/gkw955
  60. Langmead B, Trapnell C, Pop M, Salzberg SL: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009, 10(3):R25. doi:10.1186/gb-2009-10-3-r25
  61. Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, Greiner S: GeSeq - versatile and accurate annotation of organelle genomes. *Nucleic Acids Res* 2017, 45(W1):W6-w11. doi:10.1093/nar/gkx391
  62. Misra S, Harris N: Using Apollo to Browse and Edit Genome Annotations. *Current Protocols in Bioinformatics* 2005, 12(1):9.5.1-9.5.28. doi:10.1002/0471250953.bi0905s12
  63. Greiner S, Lehwark P, Bock R: OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic acids research* 2019, 47(W1):W59-w64. doi:10.1093/nar/gkz238
  64. Rice P, Longden I, Bleasby A: EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 2000, 16(6):276-277. doi:10.1016/s0168-9525(00)02024-2
  65. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R: REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic acids research* 2001, 29(22):4633-4642. doi:10.1093/nar/29.22.4633
  66. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S: MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular biology and evolution* 2013, 30(12):2725-2729. doi:10.1093/molbev/mst197
  67. Zhang D, Gao F, Jakovlic I, Zou H, Zhang J, Li WX, Wang GT: PhyloSuite: An integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. *Mol Ecol Resour* 2020, 20(1):348-355. doi:10.1111/1755-0998.13096
  68. Rozewicki J, Li S, Amada KM, Standley DM, Katoh K: MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic acids research* 2019, 47(W1):W5-w10. doi:10.1093/nar/gkz342

## Figures



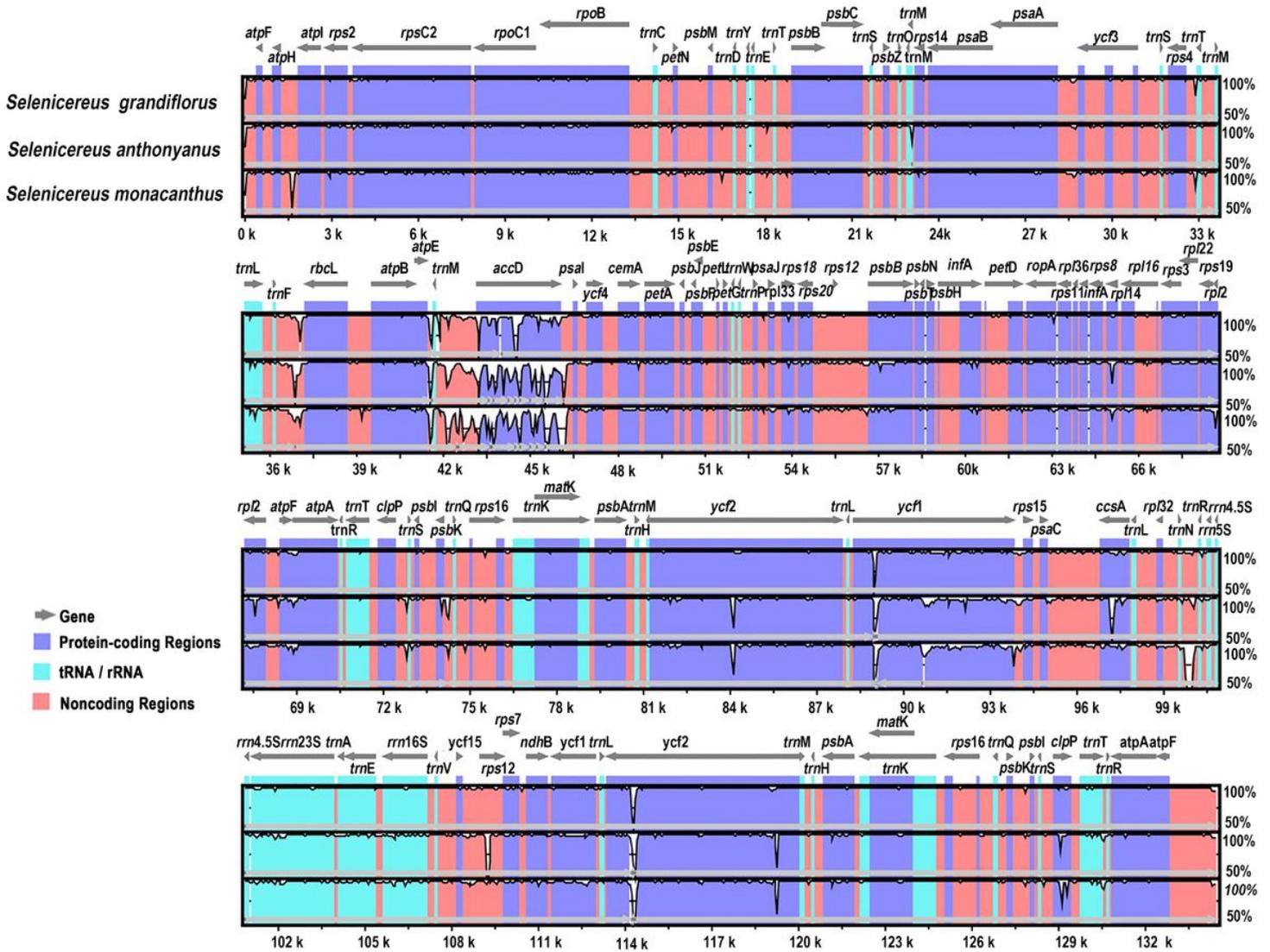
**Figure 1**

Plastid genome map of *Selenicereus* species and corresponding pictures of four plants. The thick line spacing in the inner circle represents a conservative quaternary structure, with LSC region, SSC region and a pair of IR region, and the dark gray area and light gray parts inside represent the ratio of GC and AT content, respectively. Genes inside and outside the circle were transcribed in clockwise and counterclockwise directions, respectively and genes in different functional groups have different colors.



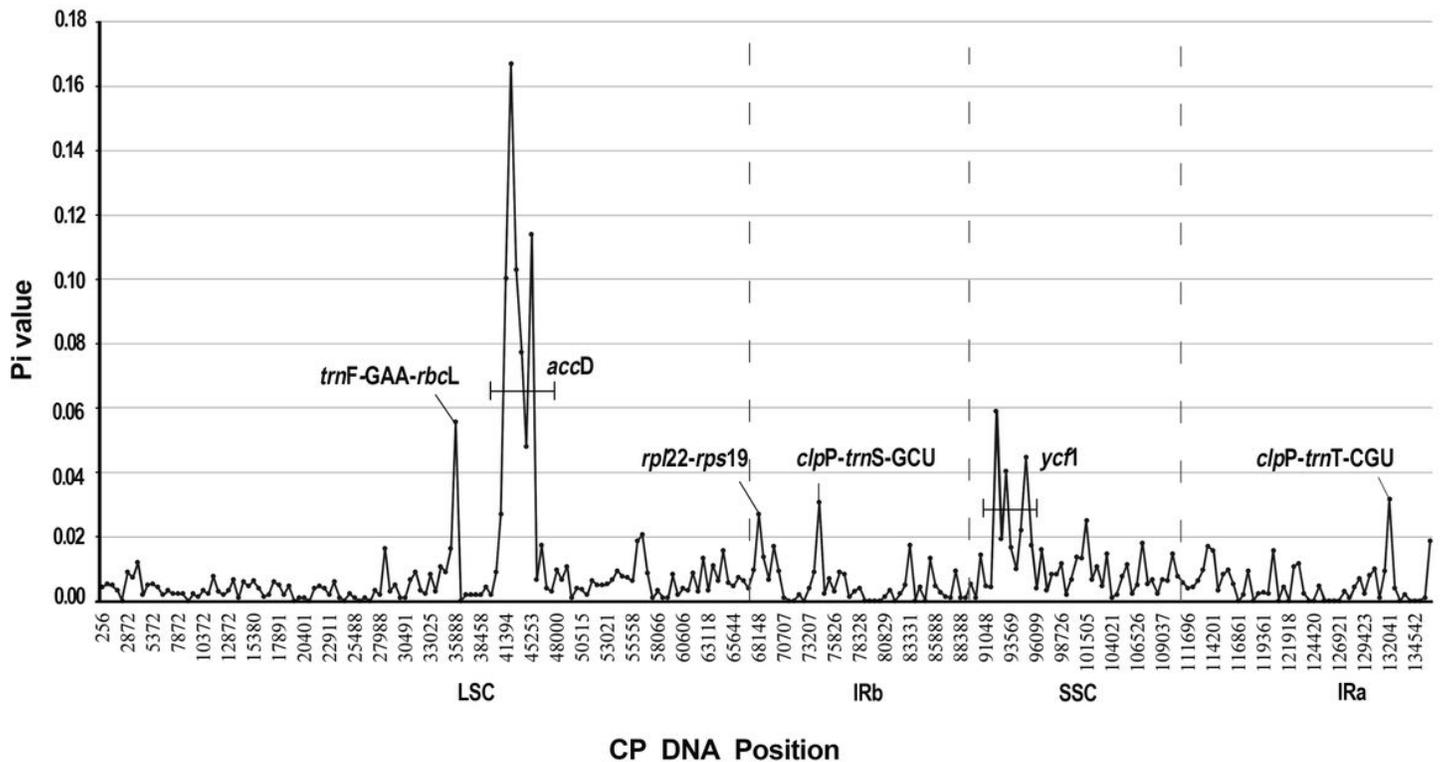
**Figure 2**

Comparison of repeated sequences in the 4 *Selenicereus* plastomes. (A). Types and numbers of repeats detected in the 4 chloroplast genomes; (B). Types and numbers of SSRs detected in the 4 plastomes; (C-F). Types and numbers of SSR motifs detected in the 4 plastomes.



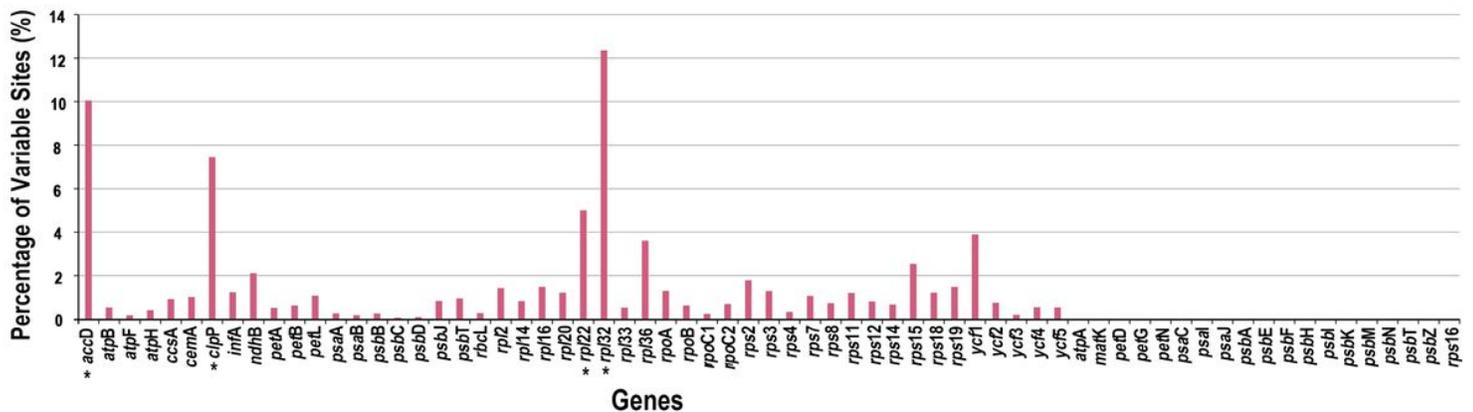
**Figure 3**

Sequence similarity of 4 *Selenicereus* species by using *S. validus* as a reference sequence and visualized in mVISTA. Different color markers represent different areas, the pink regions are conserved noncoding sequences, the purple regions are protein-coding sequences, the light blue regions are tRNA or rRNA and the gray arrows are the gene and its direction. The percentage of identity ranges from 50 to 100%, shown on the Y-axis.



**Figure 4**

The nucleotide diversity (Pi) of four *Selenicereus* plastomes. It was analyzed by using DnaSP with a sliding window analysis (window length: 500 bp, step size: 500 bp). The horizontal and vertical axes respectively represents the midpoint position of the window and the Pi value of each window. Pi values in one intergenic (*trnF-GAA-rbcL*, 0.05567) and two protein-coding genes (*accD*, 0.00667-0.167; *ycf1*, 0.004-0.059) were greater than 0.05.



**Figure 5**

Percentage of variable sites for 67 shared plastidial genes of 4 *Selenicereus* species calculated by MEGA v6.0. The four genes with the highest mutation rate have been marked with an “\*” in the figure, and they

are rpl32 (12.34%), accD (10.05%), clpP (7.44%) and rpl22 (5.00%), respectively.

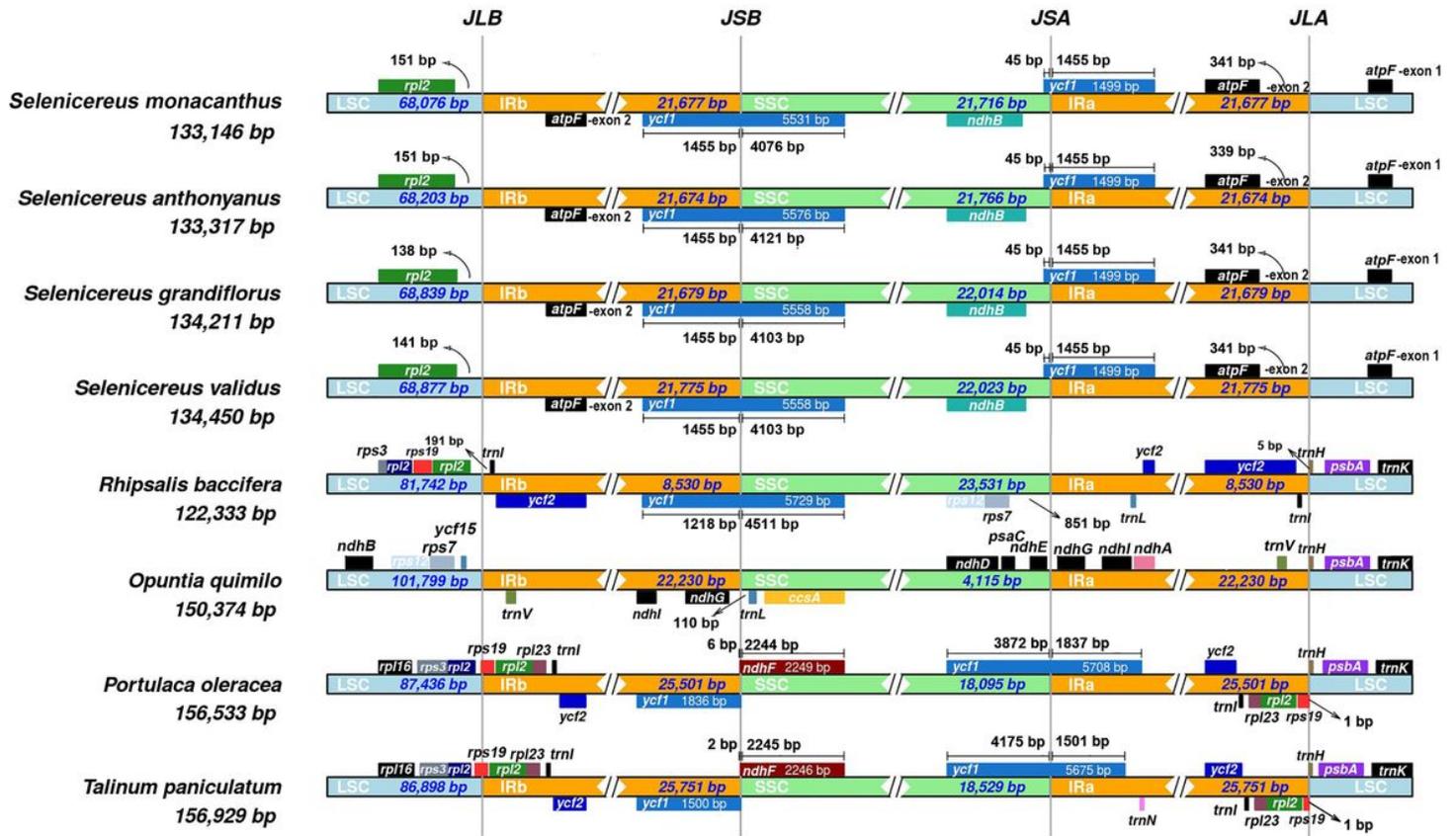


Figure 6

Comparison of borders among the LSC, SSC and IR region of 8 species. JLB, JSB, JSA and JLA represents the boundary between LSC/IRb, IRb/SSC, SSC/IRa and IRa/LSC.



- [TableS5.docx](#)
- [TableS6.docx](#)