Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Comparative and Phylogenetic Analysis of Complete Chloroplast Genomes in the Genus Hippeastrum

Yue Luo
Chinese Academy of Tropical Agricultural Sciences
Zhen Shen
Chinese Academy of Tropical Agricultural Sciences
Shuangshuang Yi
Chinese Academy of Tropical Agricultural Sciences
Liang Xi
Chinese Academy of Tropical Agricultural Sciences
Yue Liang
China Agricultural University
Wengang Yu
Hainan University
Guangsui Yang
139765728706163.com

Chinese Academy of Tropical Agricultural Sciences

Research Article

Keywords: Hippeastrum, wild species, chloroplast genome, comparative genomics, Phylogenetic Analysis

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11030845/v1

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

Additional Declarations: No competing interests reported.

Comparative and Phylogenetic Analysis of Complete Chloroplast Genomes in the Genus *Hippeastrum*

Yue Luo[1,2,3,4], Zhen Shen[1], Shuangshuang Yi[1], Liang Xi[1], Yue Liang [2,3], Wengang Yu [4], Guangsui Yang[1]*

[1]Key Laboratory of Tropical Crops Germplasm Resources Genetic Improvement and Innovation of Hainan Province, Hainan Provincial Engineering Research Center for Innovative Utilization of Tropical Ornamental Plant Germplasm, Key Laboratory of Crop Gene Resources and Germplasm Creation in South China, Ministry of Agriculture, Institute of Tropical Crop Variety Resources, Chinese Academy of Tropical Agricultural Sciences, Haikou 570100, China

[2]College of Horticulture, China Agricultural University, Beijing 100193, China

[3]Sanya Institute of China Agricultural University, Sanya 572025, China

[4]College of Tropical Agriculture and Forestry, Hainan University, Danzhou 571700, China

*** Correspondence:** Guangsui Yang, 13976572870@163.com

Abstract

Background: *Hippeastrum* is an economically important ornamental genus with high horticultural value. However, limited genomic resources have hindered the understanding of species differentiation, chloroplast genome evolution, and phylogenetic relationships within this genus. This study aimed to characterize chloroplast genome diversity and evolutionary patterns among wild *Hippeastrum* species and provide genomic resources for future systematic and breeding studies.

Results: Complete chloroplast genomes of eight wild *Hippeastrum* species were sequenced, assembled, and comparatively analyzed. The chloroplast genomes exhibited a typical quadripartite structure, with

genome sizes ranging from 157,887 to 158,116 base pairs and highly conserved gene organization containing 135-136 genes. Comparative analyses revealed strong conservation in genome structure, gene composition, and sequence arrangement among *Hippeastrum* species. Simple sequence repeat analysis showed similar repeat composition and distribution patterns across species, suggesting conserved repetitive characteristics. Selection pressure analysis indicated that most chloroplast protein-coding genes were subject to strong purifying selection, whereas several genes, including *ycf1*, *ycf2*, and *ndhD*, showed relatively accelerated evolutionary rates and represented potential evolutionary hotspots. Analysis of inverted repeat boundaries demonstrated that chloroplast genomes were highly conserved, while *Hippeastrum rutilum* exhibited distinct inverted repeat expansion and small single-copy region contraction, indicating lineage-specific structural variation. Genome-wide comparisons revealed that variable regions were mainly concentrated in non-coding regions, boundary regions, and several rapidly evolving loci. Phylogenetic analysis based on 37 chloroplast genomes resolved the major evolutionary relationships within Amaryllidaceae and supported the placement of *Hippeastrum* within Amaryllidoideae. Within *Hippeastrum*, species relationships formed three major clades, showing partial correspondence with geographic distributions and floral morphological traits.

Conclusions: This study provides comprehensive insights into chloroplast genome evolution in *Hippeastrum* and identifies potential molecular markers for species identification and phylogenetic studies. The generated chloroplast genome resources will facilitate future evolutionary research, germplasm conservation, and genetic improvement of ornamental *Hippeastrum* species.

Keywords: *Hippeastrum*, wild species, chloroplast genome, comparative genomics, Phylogenetic Analysis.

Background

The genus *Hippeastrum* (Amaryllidaceae) comprises perennial bulbous ornamental plants and is one of the most important ornamental bulb crops worldwide, characterized by large flower, diverse colors, and long flowering periods. The genus is native to tropical regions such as the Caribbean and Mexico but is mainly distributed in South America [1]. As early as 1799, the first successful interspecific hybrid was reported (*Hippeastrum reginae* × *Hippeastrum vittatum*, producing *Hippeastrum* × 'Johnsonii') [2]. In the 19th century, *Hippeastrum* was introduced into Europe and North America, where early breeding activities began. In the latter half of the 20th century, breeding became increasingly specialized, with wild diploid species being hybridized with cultivated varieties [3, 4]. With a long cultivation history, a large number of horticultural cultivars have since been developed. Today, bulb plant production has become a major global industry, spanning countries such as the United States, the Netherlands, and India. *Hippeastrum* is widely used for cut flowers, potted plants, and landscape plantings and has high ornamental and economic value [5, 6]. In addition, *Hippeastrum* bulbs are rich in various isoquinoline alkaloids, some of which exhibit antitumor, antiviral, antiparasitic, and neuroactive pharmacological activities [7], giving the genus both medicinal and scientific research value. In recent years, due to its typical bulb structure, floral organ development patterns, and hybridization advantages, *Hippeastrum* has also become an important model material for studying morphogenesis and developmental regulation in bulbous plants.

However, the taxonomy and phylogenetic relationships within the genus *Hippeastrum* remain highly controversial. The natural species are mainly distributed in South America, and there has long been no consensus on species number and delimitation within the genus [8, 9]. The situation is further complicated by extensive hybridization during breeding and cultivation, as well as the presence of polyploidy ranging from diploid to octoploid levels [10-13], along with numerous complex polyploids and aneuploid materials [14, 15]. These factors have resulted in significant morphological variation, widespread trait overlap, and mixed population structures, making it difficult for traditional identification methods based on morphological characteristics or limited

molecular markers (such as Internal transcribed spacer (ITS), *rbcL*, and *matK*) to accurately distinguish closely related taxa [16-19]. Previous studies have shown that incongruence between nuclear and chloroplast gene trees is common within Amaryllidaceae, often accompanied by deep hybridization and incomplete lineage sorting (ILS) [15], further complicating taxonomic classification and phylogenetic inference. Therefore, clarifying the taxonomic framework and evolutionary relationships within *Hippeastrum* urgently requires new molecular evidence with higher information content.

Chloroplast genomes (cpDNA) are widely used for reconstructing plant phylogeny, species identification, and DNA barcode development due to their conserved structure, uniparental inheritance, low recombination rate, and stable mutation patterns [20, 21]. *Marchantia polymorpha* [22] and *Nicotiana tabacum* [23] were among the first plants whose complete cpDNA were successfully sequenced. In recent years, advances in high-throughput sequencing technologies have greatly promoted chloroplast genomics research, and obtaining complete cpDNA sequences has become relatively convenient. Chloroplast genome sequences contain not only coding regions, non-coding regions, and RNA genes, but also important evolutionary signals such as simple sequence repeats (SSRs), dispersed repeats, highly variable intergenic regions, and boundary variations among the Large single-copy (LSC); Small single-copy (SSC), and Inverted repeats (IRs) [24-27]. Numerous studies have shown that genome-wide variation in cpDNA can significantly improve the resolution of phylogenetic analyses and can be used to develop highly stable species identification markers [28-30]. In taxa such as Amaryllidaceae, where hybridization is frequent and ploidy levels are complex, cpDNA have irreplaceable advantages in resolving maternal inheritance patterns, identifying lineage divergence, and distinguishing closely related taxa [31]. Multiple cpDNA from *Hippeastrum* species have been completely sequenced and assembled, providing important data for phylogenetic and genome evolution studies of this genus and its close relatives [32, 33]. These studies show that *Hippeastrum* cpDNA exhibits a typical quadripartite structure, consisting of a LSC,

a SSC, and two IRs. The genome size is approximately 158 kb and contains 128-129 genes, including protein-coding genes, tRNAs, and rRNAs. In addition to *Hippeastrum*, cpDNA studies of other genera in Amaryllidaceae, such as *Narcissus* and *Lycoris*, have also revealed conserved genome structures and length variation mainly caused by IR expansion and contraction, offering important evidence for intrafamilial phylogenetic relationships [34, 35]. Phylogenetic analyses based on complete cp genomes have clarified relationships within Amaryllidaceae and identified multiple evolutionary hotspot regions that can be used for species identification and DNA barcoding [36]. Overall, sequencing and comparative analyses of cpDNA not only elucidate the genomic structural characteristics and evolutionary patterns of *Hippeastrum* and other Amaryllidaceae plants, but also provide valuable references for molecular systematics, resource conservation, and genetic improvement.

However, the number of publicly available complete cpDNA of *Hippeastrum* is still very limited, with only a few species or cultivated materials reported so far [33, 37, 38], and systematic cross-species structural comparisons and variation analyses are lacking. This limitation not only hinders the construction of a robust phylogenetic framework for the genus but also restricts the development of cpDNA-based hypervariable regions, DNA barcodes, and germplasm identification markers. Therefore, constructing a comprehensive dataset of complete cpDNA covering different *Hippeastrum* materials and systematically comparing their gene composition, repeat sequence characteristics, IR expansion and contraction, distribution of highly variable regions, and overall structural variation will provide a solid foundation for taxonomic revision, phylogenetic analysis, cultivar identification, and molecular breeding in this genus.

Based on this, the present study aims to obtain complete cpDNA from eight wild *Hippeastrum* species, conduct comprehensive structural and sequence variation analyses, reveal the conservation and characteristic differences of cpDNA, and construct a high-resolution phylogenetic tree. The results of this study will provide new molecular evidence for reconstructing phylogenetic

relationships within *Hippeastrum*, offer reliable markers for species identification and germplasm evaluation, and provide important theoretical foundations and genomic resources for genetic improvement, elite parent selection, and chloroplast engineering in Amaryllidaceae plants.

Results

Chloroplast Genome Features

After sequencing and data filtering of the cpDNA from eight wild *Hippeastrum* species, the average amount of clean data per sample was approximately 6 Gb. The Q20 values (percentage of bases with a quality score ≥ 20) were all above 98.69%, and the Q30 values (percentage of bases with a quality score ≥ 30) exceeded 95.27%, with GC content ranging from 36.77% to 39.05% (Table S1). These results indicate that the sequencing data were of high quality and suitable for subsequent genome assembly and comparative analyses.

Following genome assembly and annotation, the complete chloroplast genome information of the eight *Hippeastrum* species was obtained (Table 1). The genome sizes ranged from 157,887 bp (*H. papilio*) to 158,116 bp (*H. psittacinum*), with GC content varying between 37.90% and 37.96%. All eight cpDNA exhibited a typical quadripartite structure, consisting of a pair of inverted repeat regions (IRs), a large single-copy region (LSC), and a small single-copy region (SSC) (Figure 1). The length of the LSC region ranged from 86,087 to 86,515 bp, the SSC region from 18,238 to 18,285 bp, and the IR regions from 26,674 to 26,840 bp. Gene content and arrangement were highly conserved among the eight genomes. Each genome encoded 135–136 unique genes, including 85–86 protein-coding genes, 8 rRNA genes, and 42 tRNA genes. Several genes located within the IR regions (e.g., *ndhB*, *rpl2*, *rpl23*, *rps12*, *rps19*, *rps7*, and *ycf2*) were present in two copies. Protein-coding genes accounted for approximately 63% of the total gene number, while rRNA and tRNA genes accounted for approximately 5.9% and 31%, respectively.

Comparative analysis of the cpDNA of the eight *Hippeastrum* species revealed a highly conserved gene composition based on functional classification. A total of 66 core genes were identified across all species, covering multiple functional categories related to photosynthesis, self-replication, and other biological processes (Table 2). Specifically, these shared genes included subunits of ATP synthase, photosystem II, NADH dehydrogenase, cytochrome b/f complex, and photosystem I, as well as genes encoding ribosomal large and small subunit proteins, RNA polymerase, acetyl-CoA carboxylase (*accD*), cytochrome c synthesis protein (*ccsA*), envelope membrane protein (*cemA*), protease (*clpP*), translation initiation factor (*infA*), and maturase (*matK*). In addition, all species contained several conserved open reading frames with unknown functions (*ycf1*, *ycf2*, and *ycf4*).

The complete chloroplast genome sequences of these *Hippeastrum* species have been deposited in GenBank under accession numbers PX724908, PX696961, PX724625, PX724626, PX724628, PX724629, PX724630, and PX724631.

Table 1 Feature of complete chloroplast genomes of *Hippeastrum.*

Species name	***H. evansiae***	***H. aulicum***	***H. papilio***	***H. haywardii***	***H. correiense***	***H. curitibanum***	***H. machupijchense***	***H. psittacinum***
Length	158077	158101	157887	158081	158113	158113	158072	158116
GC(%)	37.92	37.9	37.96	37.93	37.94	37.94	37.94	37.94
accession number	PX724908	PX696961	PX724625	PX724626	PX724628	PX724629	PX724630	PX724631
LSC length(bp)	86173	86515	86087	86179	86155	86155	86153	86158

SSC length(bp)	18272	18238	18282	18270	18278	18278	18285	18280
IR length(bp)	26816	26674	26759	26816	26840	26840	26817	26839
Gene number	135	136	135	135	136	136	136	136
Protein-coding gene number	85	86	85	85	86	86	86	86
Protein-coding gene(%)	62.96	63.24	62.96	62.96	63.24	63.24	63.24	63.24
rRNA gene number	8	8	8	8	8	8	8	8
rRNA(%)	5.93	5.88	5.93	5.93	5.88	5.88	5.88	5.88
tRNA gene number	42	42	42	42	42	42	42	42
tRNA(%)	31.11	30.88	31.11	31.11	30.88	30.88	30.88	30.88

Table 2 Gene composition of the chloroplast genomes of *Hippeastrum.*

Category for genes	Group of gene	Name of gene

Genes for photosynthesis	ATP synthase	*atpA*, *atpB*, *atpF*, *atpH*, *atpI*
	NADH-dehydrogenase	*ndhA, ndhB, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK*
	Cytochrome b/f complex	*petA, petB, petD, petG, petL, petN*
	Photosystem I	*psaA, psaB, psaC, psaI, psaJ*
	Photosystem II	*psbA, psbB, psbC, psbD, psbE, psbF, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ, ycf3*
	Rubisco	*rbcL*
Self replication	Large subunit of ribosome	*rpl14, rpl16, rpl2, rpl20, rpl22, rpl23, rpl32, rpl33, rpl36*
	DNA-dependent RNA polymerase	*rpoA, rpoB, rpoC1, rpoC2*
	Small subunit of ribosome	*rps11, rps12, rps14, rps15, rps16, rps18, rps19, rps2, rps3, rps4, rps7, rps8*
Other genes	Subunit of Acetyl-CoA-carboxylase	*accD*
	C-type cytochrome synthesis gene	*ccsA*
	Envelope membrane protein	*cemA*
	Protease	*clpP*
	Translational initiation factor	*infA*
	Maturase	*matK*
Genes of unknown function	Conserved open reading frames	*ycf1, ycf2, ycf4*

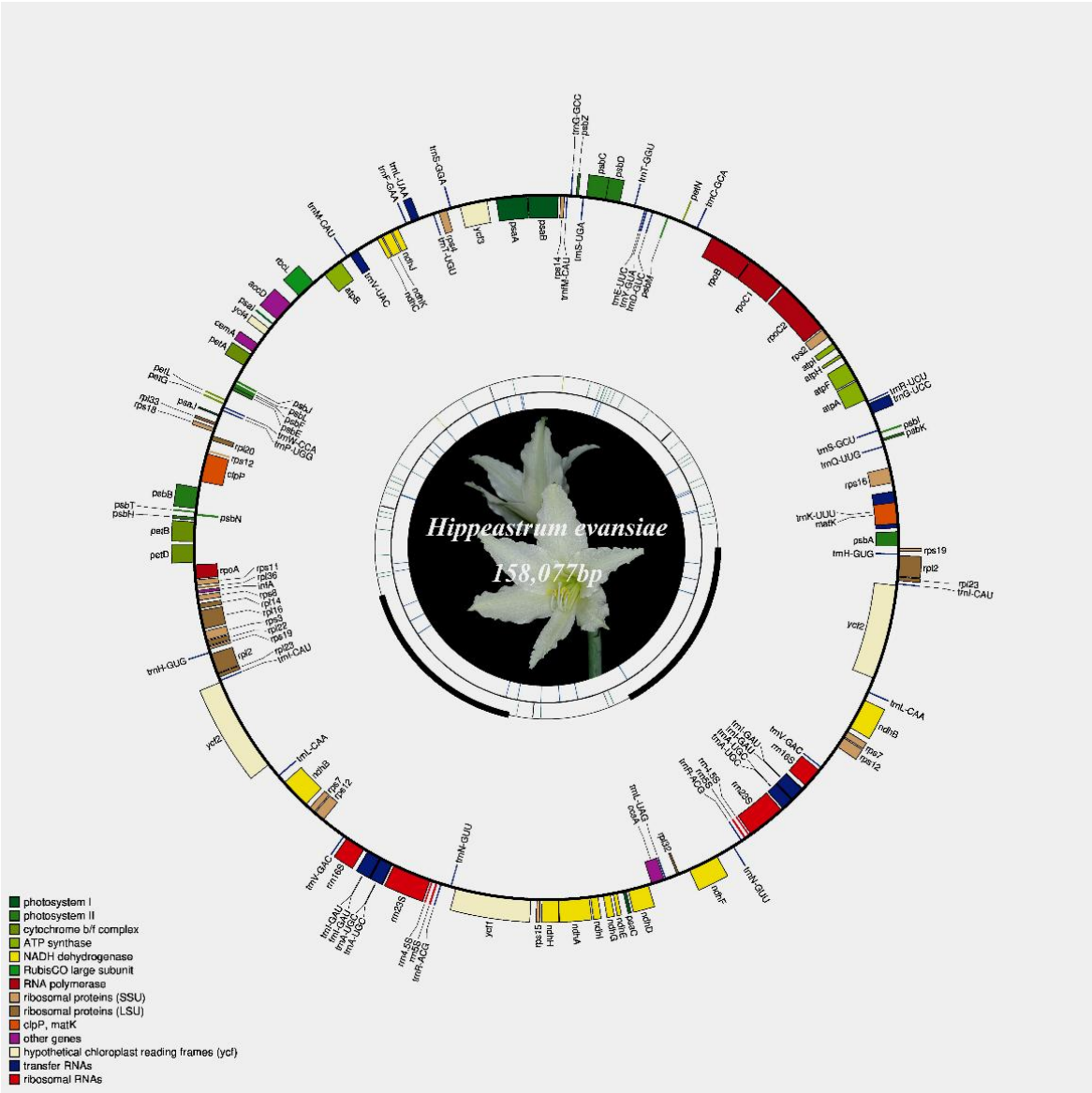


Figure 1 Gene maps of *Hippeastrum evansiae.*

Codon usage and nucleotide diversity

The protein-coding codon usage patterns were highly similar among the eight *Hippeastrum* species. Codons with the highest RSCU values were ATT, AAA, GAA, TTT, and AAT, whereas those with the lowest RSCU values were TGA, TAG, TAA, TGC, and CGC. (Table 3).

Relative synonymous codon usage (RSCU) was employed as an intuitive measure to evaluate codon usage bias. Codon preference was classified into four categories: high preference (RSCU >

1.3), moderate preference (1.2 < RSCU < 1.3), low preference (1.0 < RSCU < 1.2), and no preference (RSCU = 1.0). In the chloroplast protein-coding genes of the eight *Hippeastrum* species, 64 codons encoding 20 amino acids were identified. Except for cysteine, codons corresponding to all other amino acids exhibited a high preference (Table S2).

Table 3 Codon usage for *H. evansiae* chloroplast genome.

Amino Acid	Codon	Number	RSCU	Amino Acid	Codon	Number	RSCU
Ala	GCT	623	2.40	Gln	CAA	673	2.59
	GCA	399	1.53		CAG	225	0.87
	GCG	158	0.61	Ser	TCA	422	1.62
	GCC	219	0.84		TCC	364	1.40
Val	GTC	189	0.73		TCG	205	0.79
	GTT	502	1.93		TCT	558	2.15
	GTA	516	1.98		AGC	110	0.42
	GTG	203	0.78		AGT	404	1.55
Ile	ATT	1055	4.06	Thr	ACA	387	1.49
	ATA	703	2.70		ACC	246	0.95
	ATC	470	1.81		ACG	156	0.60
Lys	AAA	1011	3.89		ACT	535	2.06
	AAG	374	1.44	Arg	AGA	512	1.97
Glu	GAA	998	3.84		AGG	165	0.63
	GAG	364	1.40		CGA	346	1.33
Phe	TTT	919	3.53		CGC	102	0.39
	TTC	567	2.18		CGG	140	0.54
Asn	AAT	917	3.53		CGT	361	1.39
	AAC	298	1.15	His	CAC	164	0.63

Asp	GAT	860	3.31	Trp	CAT	478	1.84
	GAC	213	0.82		TGG	448	1.72
Leu	TTA	790	3.04	Pro	CCA	305	1.17
	TTG	572	2.20		CCC	249	0.96
	CTT	557	2.14		CCG	138	0.53
	CTA	384	1.48		CCT	386	1.48
	CTC	186	0.72	Cys	TGC	72	0.28
	CTG	163	0.63		TGT	228	0.88
Tyr	TAT	758	2.91	Met	ATG	604	2.32
	TAC	206	0.79	Stop	TAA	38	0.15
Gly	GGA	725	2.79		TAG	26	0.10
	GGC	168	0.65		TGA	21	0.08
	GGG	308	1.18				
	GGT	568	2.18				

Repeat Sequence Analysis (SSR)

SSR analysis of the eight *Hippeastrum* species revealed variation in their distribution patterns. The total number of SSRs and compound SSRs among the eight species ranged from 37 to 48 and from 3 to 7, respectively, within a relatively narrow range, indicating a high level of conservation (Figure 2 and Table S3). Mononucleotide repeats composed of A/T were the most common type. Among all SSRs identified in the eight cpDNA, mononucleotide repeats accounted for the highest proportion, exceeding 82%, whereas dinucleotide and trinucleotide repeats were relatively rare, with only 1 to 3 detected in each genome. (Table 4). Owing to their high abundance, conserved distribution, and stable repeat types, these SSRs may have potential applications in species identification, phylogenetic studies, and molecular marker development. However, their practical utility requires further validation across a broader range of samples.

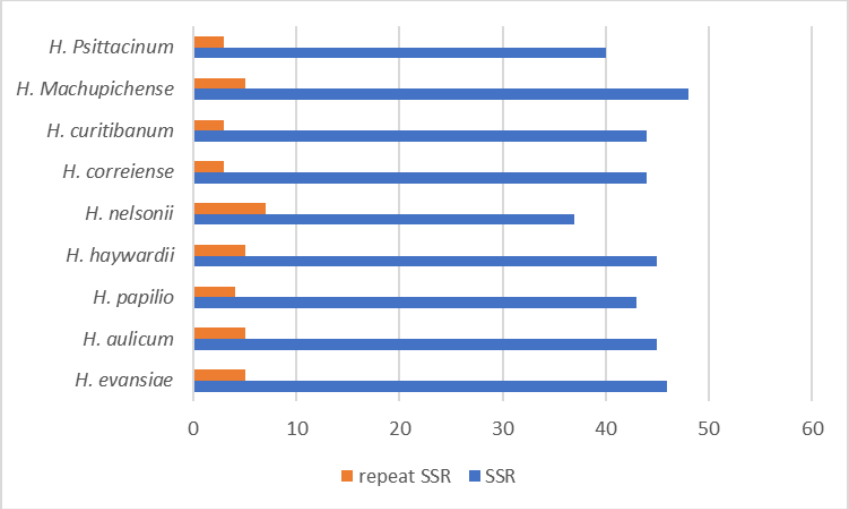


Figure 2 Simple sequence repeats (SSRs) in chloroplast genomes of *Hippeastrum.*

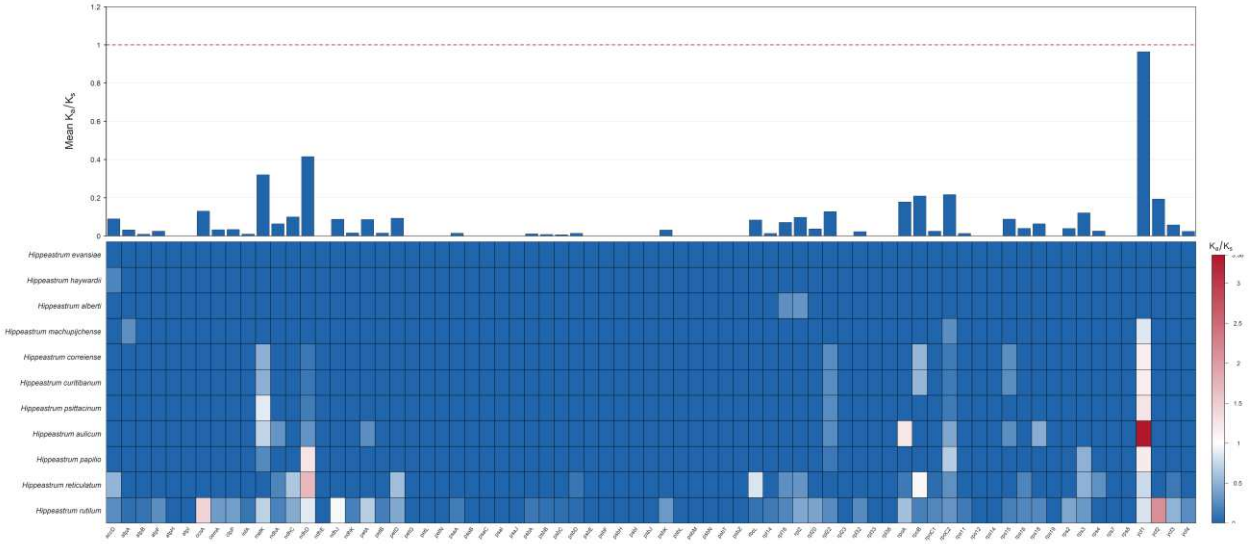
Table 4 Distribution of SSR repeat types in the chloroplast.

Species	Total SSRs	Mono	Di	Tri	Compound
H. evansiae	41	34(83% A/T)	2(TA)	1(AAT)	4
H. haywardii	40	33(83% A/T/C)	2(TA)	1(ATT)	4
H. machupijchense	43	36(84% A/T/C)	2(TA)	1(AAT)	4
H. correiense	41	36(88% A/T/C)	2(TA)	1(AAT)	2
H. curitibanum	41	36(88% A/T/C)	2(TA)	1(ATT)	2
H. psittacinum	37	32(86% A/T)	2(TA)	1(ATT)	2
H. aulicum	40	33(83% A/T/C)	2(TA)	1(ATT)	4
H. papilio	39	32(82% A/T)	3(AT/TA)	1(AAT)	3

Selection Pressure Analysis (Ka/Ks)

Selective pressures on chloroplast protein-coding genes were assessed by estimating the ratios of nonsynonymous substitution rate (Ka) to synonymous substitution rate (Ks) (Ka/Ks) through pairwise comparisons between *H. vittatum* (used as the reference species) and the other 11 retained *Hippeastrum* species. Most protein-coding genes showed Ka/Ks values substantially below 1 across the comparisons analyzed, indicating pervasive purifying selection during *Hippeastrum* evolution (Figure 3 and Table S4). This overall pattern suggests that genes involved in essential chloroplast functions, including photosynthesis, transcription, translation, and energy metabolism, have been maintained under strong functional constraints.

However, Ka/Ks values varied among genes and lineages. *ycf1* exhibited the highest Ka/Ks ratio, with a particularly elevated value in *H. aulicum*, suggesting lineage-specific acceleration of nonsynonymous sequence evolution. *ycf2* also showed a comparatively high ratio in *H. rutilum* and therefore represents another candidate rapidly evolving chloroplast gene. Elevated ratios were also detected for *ccsA*, *ndhD*, and *rpoA* in individual species comparisons, further supporting heterogeneous evolutionary rates among chloroplast genes. These elevated values identify candidate loci for further investigation; however, additional branch- or site-specific analyses are required to determine whether they reflect positive selection.



209

210 **Figure 3** Ka/Ks ratios of chloroplast protein-coding genes in 12 *Hippeastrum* species using *H.*
211 *vittatum* as the reference species. (A) Bar chart showing the average Ka/Ks values of 73 chloroplast
212 CDS sequences calculated from pairwise comparisons between *H. vittatum* and the other 11
213 *Hippeastrum* species. Blue and orange indicate genes under purifying selection (Ka/Ks < 1) and
214 putative positive selection (Ka/Ks > 1), respectively, with the dashed line representing Ka/Ks = 1. (B)
215 Heatmap showing Ka/Ks values of 73 CDS sequences across 12 *Hippeastrum* species, normalized
216 using simple linear scaling.

217

218 **IR Boundary Analysis**

219 The IR/SC boundaries of the cpDNA of 12 *Hippeastrum* species were compared using IRscope
220 (Figure 4). The results showed that, except for *H. rutilum*, the boundary structures and arrangements
221 of adjacent genes were highly conserved among the other 11 species. The complete cpDNA ranged
222 from 157,887 to 162,215 bp in length, with the LSC, SSC, and individual IR regions ranging from
223 86,088 to 86,933 bp, 5,418 to 18,409 bp, and 26,673 to 34,932 bp, respectively. Relatively little
224 interspecific variation was observed at the JLB and JLA boundaries, which were mainly associated
225 with the *rpl22-rps19* and *rpl2-rps19-trnH-psbA* gene clusters, respectively. In contrast, most

boundary variation occurred at the JSB and JSA junctions and involved differences in the positions and boundary-spanning lengths of *ycf1* and *ndhF*. In most species, the complete *ycf1* gene spanned the JSA boundary, while a partial duplicated copy of approximately 981-996 bp occurred in the opposite IR region, indicating minor expansion and contraction of the IR regions within the genus.

Compared with the other species, *H. rutilum* exhibited the most distinct boundary structure. Its individual IR region expanded to 34,932 bp, whereas its SSC region contracted to only 5,418 bp. The *rpl32* gene was located near the JSB boundary, while *ndhE*, *ndhG*, and *ndhA* occurred near the JSA boundary; the *ndhA* gene was 1,813 bp in length. This arrangement differed markedly from the pattern observed in the other species, whose JSB and JSA boundaries were mainly associated with *ycf1* and *ndhF*. These results indicate that substantial expansion of the IR region toward the SSC region in *H. rutilum* incorporated several genes originally located within the SSC into the IR region or repositioned them near the IR/SSC junctions. Therefore, *H. rutilum* experienced pronounced IR expansion and SSC contraction and displayed the greatest degree of chloroplast genome boundary divergence among the 12 species.

Overall, the IR regions of *Hippeastrum* cpDNA are highly conserved, whereas *H. rutilum* represents an exceptional lineage exhibiting pronounced IR expansion and SSC contraction. This structural divergence highlights the dynamic nature of IR boundary evolution and suggests that IR expansion and contraction may contribute to chloroplast genome differentiation among *Hippeastrum* species.

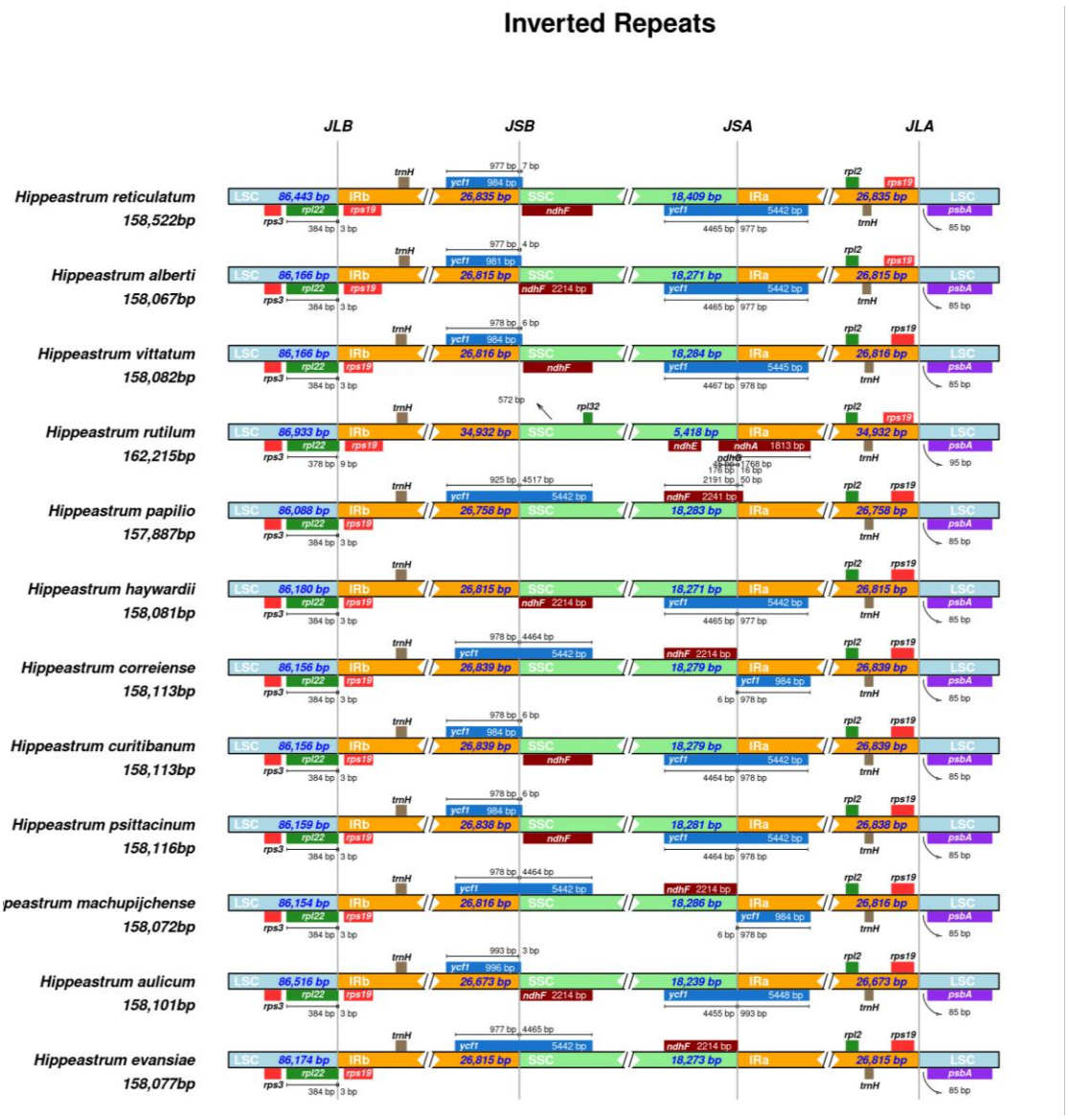


Figure 4 Comparison of the LSC, SSC, and IR regions among twelve cpDNA.

Chloroplast genome-wide alignment analysis

The Comparative analysis based on mVISTA revealed that cpDNA within the genus *Hippeastrum* exhibited high sequence conservation and strong collinearity. Using *H. evansiae* as the reference genome, the 12 *Hippeastrum* cpDNA showed high sequence similarity across most coding regions (Figure 5A). Highly conserved regions were mainly observed in several functional gene clusters, including *rpoC2-rpoC1-rpoB*, *rbcL*, *psbB-psbT-psbN-petB-petD*, ribosomal protein gene clusters

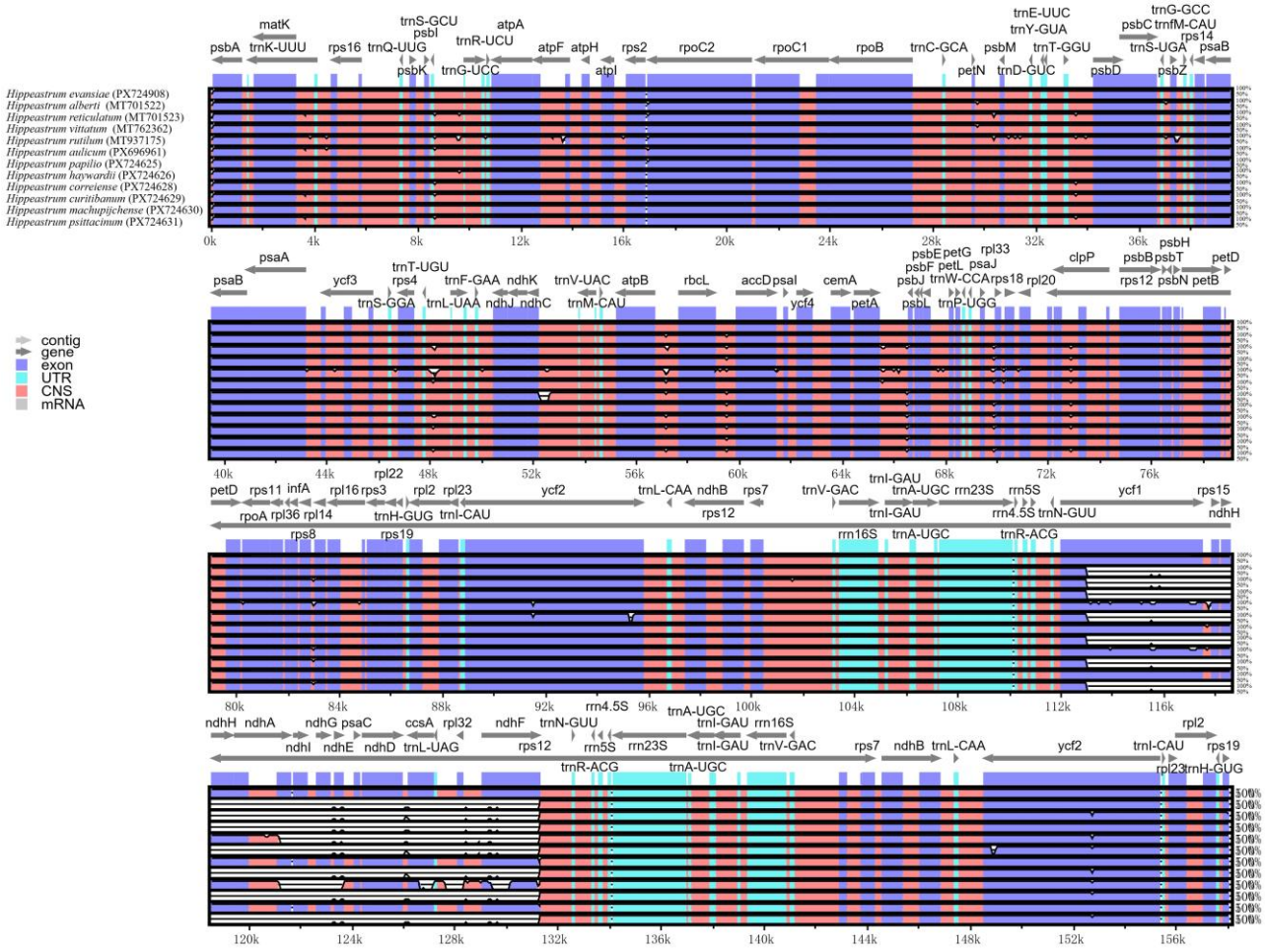
(*rpl*/*rps*), and ribosomal RNA genes (*rrn16S*, *rrn23S*, *rrn5S*, and *rrn4.5S*). In contrast, non-coding regions and some boundary regions exhibited relatively reduced sequence similarity, indicating that these regions accumulated greater sequence variation among *Hippeastrum* species. The most variable regions within *Hippeastrum* were mainly concentrated in several intergenic spacer regions and selected coding regions, including *trnK-UUU*/*matK*, *rps16-trnQ-UUG*, *trnS-GCU*/*psbI*/*trnG-UCC*, *atpF-atpH*, *petN-psbM*, *psbM-trnD-GUC*, *trnT-GGU*/*trnE-UUC*/*trnY-GUA*, *rpl32-trnL*, *ndhF*, and regions surrounding *ycf1*. These divergent regions may represent potential candidate loci for species identification, phylogenetic analyses, and chloroplast molecular marker development within *Hippeastrum*.

To further evaluate chloroplast genome variation at the family level, representative species from several genera of Amaryllidaceae were additionally compared with *H. evansiae* using mVISTA (Figure 5B). The results showed that cpDNA among different genera within Amaryllidaceae also maintained relatively high overall conservation, particularly in protein-coding genes and rRNA regions, indicating conserved genome organization and gene composition throughout the family. However, compared with the within-genus comparison, intergeneric comparisons exhibited more extensive reductions in sequence similarity and greater sequence divergence, suggesting that chloroplast genome differentiation generally increased with taxonomic distance. In comparisons among different genera, highly variable regions were mainly distributed in non-coding regions, untranslated regions (UTRs), single-copy regions, and regions near the IR/SC boundaries. These regions included *rps16-trnQ*, *trnS-psbI-trnG*, *atpF-atpH*, *petN-psbM*, *trnT*/*trnE*/*trnY*, *rpl32-trnL*, *ccsA-ndhD*, *ndhF*, *ycf1*, and regions associated with the IR/SSC boundaries. Among these regions, *ycf1* and *ndhF* displayed relatively high sequence divergence in both intra- and intergeneric comparisons, suggesting that these loci may represent rapidly evolving regions within Amaryllidaceae cpDNA.

Nucleotide diversity analysis based on the complete cpDNA of 12 *Hippeastrum* species revealed a generally low level of sequence variation, with most regions showing Pi values below 0.005 (Figure 6). However, several highly variable regions were identified, including the *petA-psbJ*, *psbJ-psbL*, *ccsA-ndhD*, *ndhD*, and *ycf1* regions, which exhibited relatively higher Pi values. Gene-based analysis further showed that *trnL-UAA* had the highest nucleotide diversity (Pi = 0.05848), followed by *ndhD* (Pi = 0.00978), whereas most protein-coding and RNA genes displayed low nucleotide diversity. These results indicate that sequence variation in *Hippeastrum* cpDNA is mainly concentrated in a limited number of rapidly evolving regions, while most functional regions remain highly conserved.

Overall, the mVISTA and nucleotide diversity analyses together revealed a conserved yet heterogeneous evolutionary pattern in *Hippeastrum* cpDNA. The high conservation observed in coding and rRNA regions reflects strong functional constraints maintaining essential chloroplast functions, whereas the elevated sequence divergence detected in non-coding regions, IR/SC boundary regions, and several loci such as *ycf1* and *ndhD* indicates the presence of lineage-specific evolutionary hotspots. These highly variable regions may contribute to chloroplast genome differentiation among *Hippeastrum* species and represent potential candidate markers for species identification, phylogenetic reconstruction, and evolutionary studies.

A



304

305

306

307

308

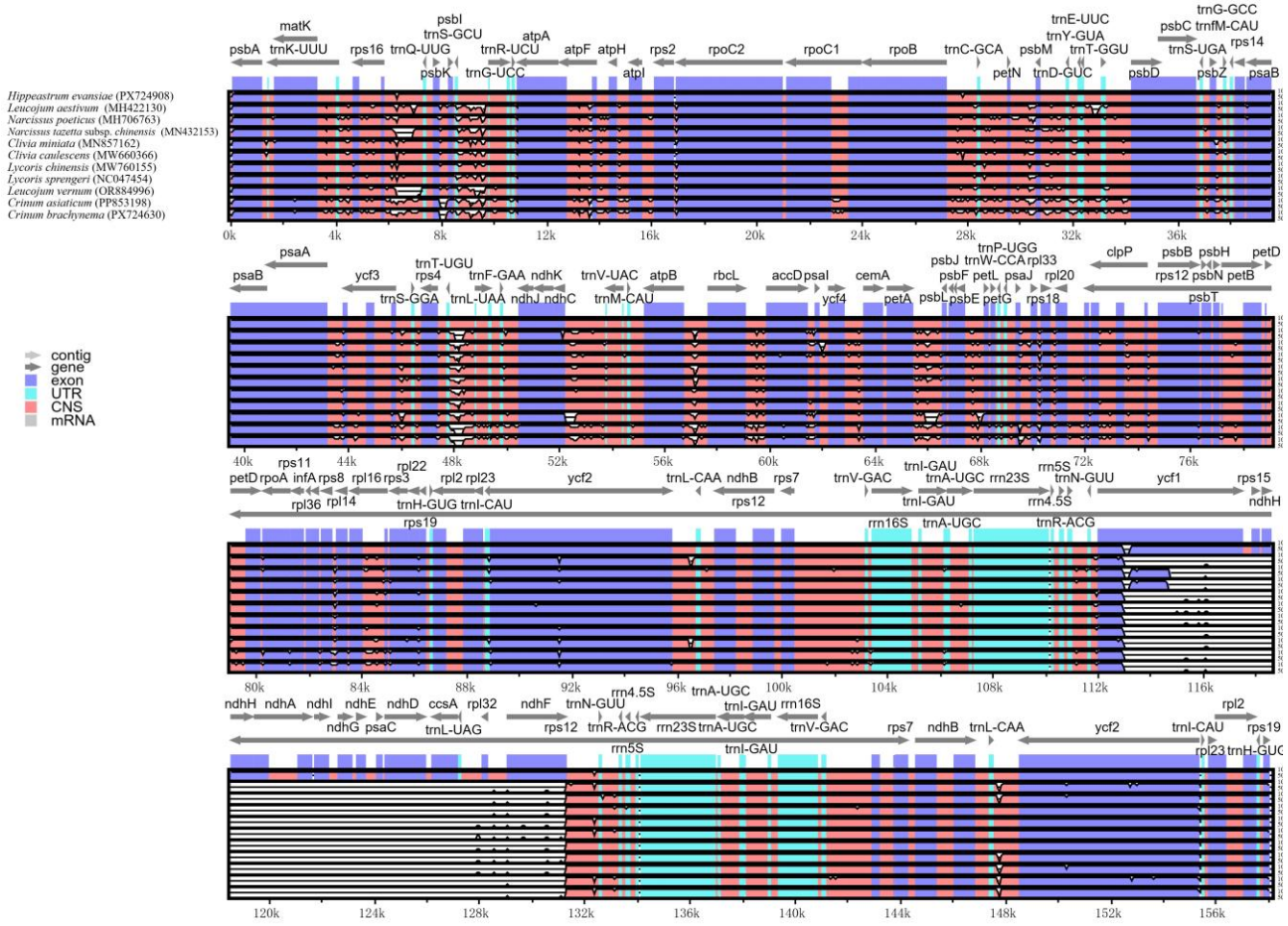
309

310

311

312

B



313 **Figure 5** Visualization of the chloroplast genome comparison of 22 species using *H. evansiae* as a
314 reference. (A) Sequence comparison of 12 *Hippeastrum* cpDNA using *H. evansiae* as the reference
315 genome. (B) Sequence comparison of representative cpDNA from different genera of
316 Amaryllidaceae using *H. evansiae* as the reference genome. The horizontal axis represents the
317 coordinates within the chloroplast genome, and the vertical axis indicates the percentage identity,
318 ranging from 50 % to 100 %. The colors represent different regions: blue for exons, green for introns,
319 and red for intergenic regions.
320
321

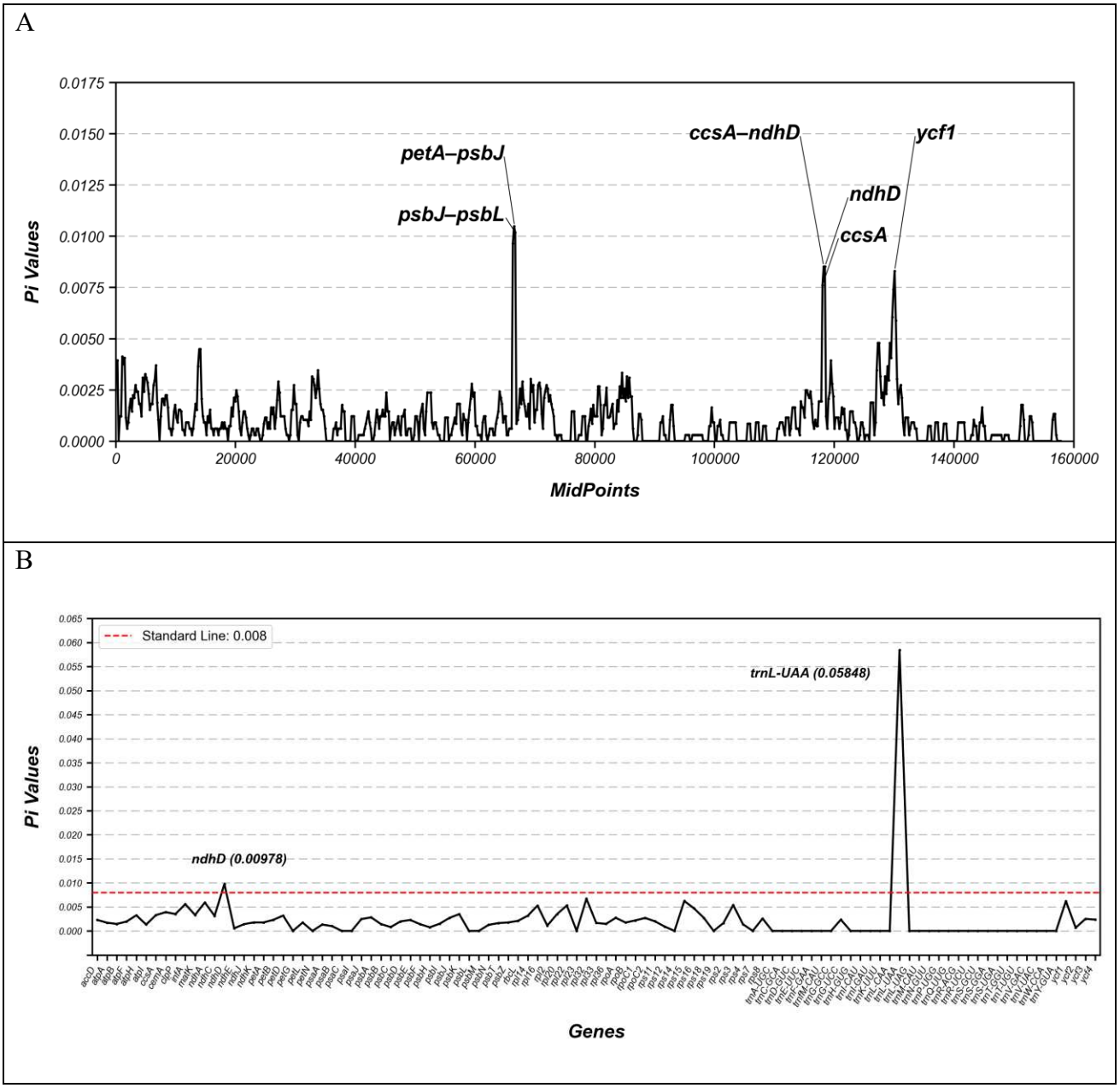


Figure 6 Nucleic acid variation in matching areas of 12 cpDNA of *Hippeastrum*.

Phylogenetic Analysis

Based on maximum likelihood phylogenetic analysis of 37 cpDNA representing Amaryllidaceae and related lineages (including an outgroup), the family was resolved into three well-supported major clades corresponding to Agapanthoideae, Allioideae, and Amaryllidoideae (Figure 7A). Within Amaryllidoideae, three well-supported clades (Clade A-C) were recovered, reflecting distinct

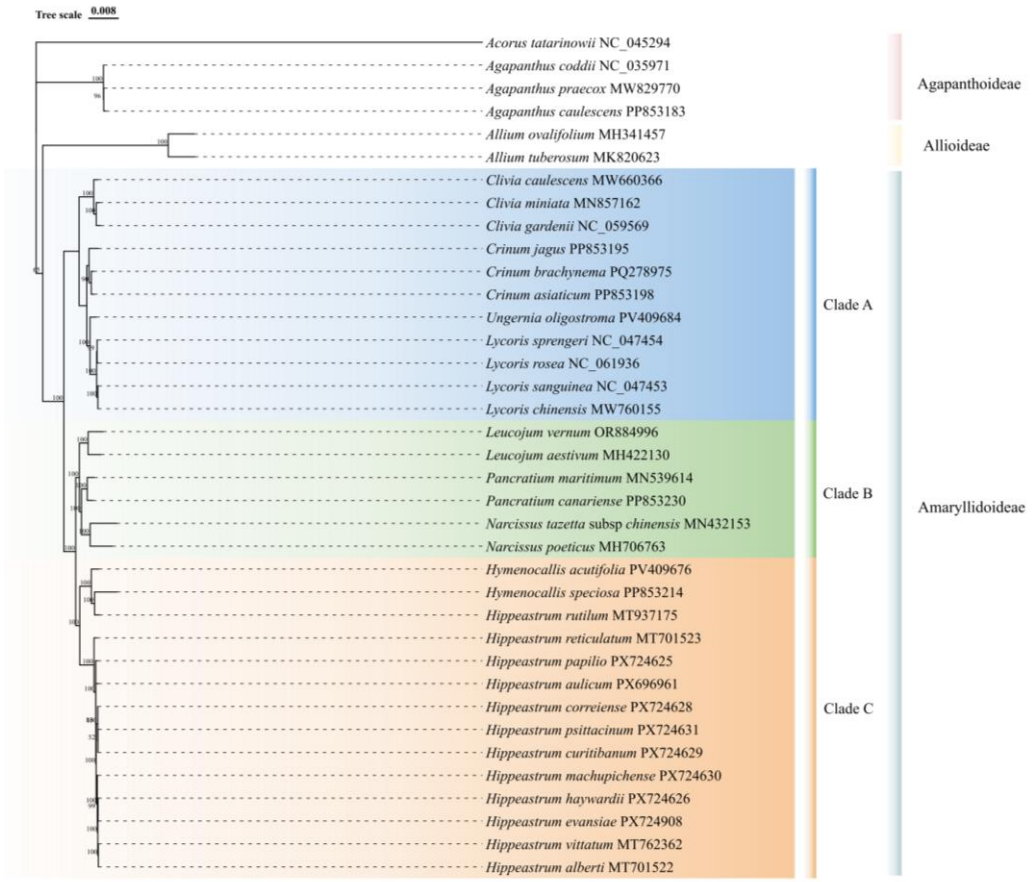
329 evolutionary lineages. Clade A comprised *Crinum*, *Clivia*, *Lycoris*, and *Ungernia* species,
330 representing relatively early-diverging lineages characterized by conserved bulbous morphology and
331 typical umbellate inflorescences. Clade B included *Leucojum*, *Narcissus*, and *Pancratium*, which
332 exhibited more specialized floral structures, including reduced floral size and derived morphological
333 traits such as the corona in *Narcissus*. Clade C contained *Hymenocallis* and *Hippeastrum*,
334 representing a more recently diverged lineage.

335 The results indicated that *Hippeastrum* species could be divided into three major clades, with
336 divergence patterns potentially corresponding to geographic distribution in their native ranges and
337 floral morphological traits (Figure 7B). Most nodes received high bootstrap support (>89%),
338 suggesting that the phylogenetic relationships are generally reliable. However, some nodes showed
339 relatively low support, such as the branch between *H. machupijchense* and *H. alberti*, which had a
340 bootstrap value of 51%. The first clade included *H. evansiae*, *H. haywardii*, *H. machupijchense*, *H.*
341 *alberti*, and *H. vittatum*. These species are mainly distributed in the Andes Mountains and adjacent
342 regions, including Bolivia, Argentina, and the Brazilian Plateau. Most species in this clade are
343 terrestrial and are characterized by relatively narrow tepals, smaller stigmas, and relatively small
344 flowers; The second clade consisted of *H. correiense*, *H. curitibanum*, and *H. psittacinum*, which are
345 mainly distributed in the Brazilian Plateau and surrounding hilly regions. These species are
346 characterized by broader tepals, relatively larger flowers, and medium-sized stigmas; The third clade
347 included *H. papilio*, *H. aulicum*, *H. reticulatum*, and *H. rutilum*, which are primarily distributed in
348 the Atlantic Forest region of southeastern Brazil. Among them, *H. papilio* and *H. aulicum* are
349 lithophytic or epiphytic species and exhibit strong ecological adaptability. Species in this clade
350 generally have broader tepals, and some species exhibit notched inner tepals and relatively larger
351 stigmas.

352 Overall, the phylogenetic branching patterns of *Hippeastrum* appear to correspond, to some
353 extent, with geographic distribution and certain floral traits. However, due to the presence of low-

support nodes and the limited sample size, these results provide only a preliminary reference for understanding phylogenetic relationships among *Hippeastrum* species.

A



B

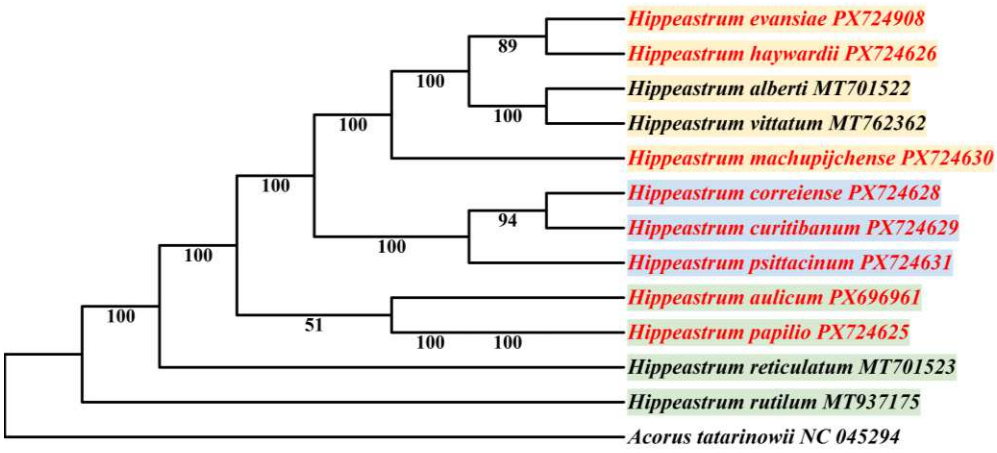


Figure 7 Phylogenetic analysis. (A) Phylogenetic tree inferred from 37 representative chloroplast genomes across Amaryllidaceae and related lineages. (B) Phylogenetic tree focusing on *Hippeastrum* and closely related taxa based on 13 chloroplast genomes. *Acorus tatarinowii* was used as the outgroup to root the phylogenetic trees. Numbers at nodes indicate bootstrap support values from maximum likelihood analysis.

Discussion

Chloroplast genomes generally exhibit highly conserved evolutionary patterns due to their essential roles in photosynthesis, energy metabolism, and plastid genetic regulation. In this study, the newly assembled cpDNA of *Hippeastrum* species, together with comparative analyses of additional congeneric and related taxa, revealed a characteristic evolutionary pattern in which overall genome structure was highly conserved, whereas specific genomic regions exhibited lineage-specific divergence. The conserved genome size, GC content, gene composition, and gene arrangement observed among *Hippeastrum* species indicate that chloroplast genome evolution within this genus

371 has been primarily constrained by functional requirements [33, 37-39]. Similar patterns have been
372 reported in other angiosperm lineages [40], where plastid genomes maintain stable organization
373 because most retained genes are involved in essential biological processes, including photosynthesis,
374 transcription, and translation. Therefore, the structural conservation observed in *Hippeastrum* cpDNA
375 likely reflects strong functional constraints acting on plastid genomes during evolution.

376 Although the overall architecture of *Hippeastrum* cpDNA was highly conserved, evolutionary
377 divergence was not uniformly distributed across the genome. Codon usage analysis revealed a strong
378 preference for A/T-ending codons, which is a common feature of angiosperm cpDNA and may reflect
379 compositional constraints and translational selection acting on plastid genes [41-43]. In *Hippeastrum*,
380 all codons except those encoding cysteine showed high preference, whereas in *Camellia*, only about
381 70% of preferred codons showed high preference [40]. This result further suggests that the cpDNA of
382 *Hippeastrum* are relatively conserved, as strong codon usage bias is a common feature in higher plant
383 cpDNA. Meanwhile, the conserved distribution pattern of SSRs among species further supports the
384 stability of cpDNA within *Hippeastrum*. Previous research has shown that chloroplast SSRs, due to
385 maternal inheritance, low recombination rate, and relatively stable mutation patterns, are suitable for
386 species identification, population genetic structure analysis, and phylogeographic studies [44, 45].
387 Therefore, the chloroplast SSR loci identified in this study may serve as potential molecular markers
388 for future germplasm evaluation, phylogeography, and phylogenetic studies in *Hippeastrum*.

389 Despite the overall conservation of chloroplast genome organization, selective pressure analysis
390 revealed heterogeneous evolutionary rates among different protein-coding genes. The majority of
391 chloroplast genes showed Ka/Ks ratios substantially lower than 1, indicating that strong purifying
392 selection has maintained the functional integrity of genes involved in essential plastid processes. This
393 pattern suggests that most chloroplast genes are subject to strong evolutionary constraints and have
394 retained conserved functions throughout *Hippeastrum* diversification. However, several genes,
395 including *ycf1*, *ycf2*, *ccsA*, *ndhD*, and *rpoA*, exhibited relatively elevated Ka/Ks values in specific

lineages, suggesting that these loci represent regions with accelerated sequence evolution. Notably, *ycf1* showed the highest evolutionary rate in *H. aulicum*, whereas *ycf2* displayed relatively increased Ka/Ks values in *H. rutilum*. The elevated Ka/Ks ratios should not be interpreted directly as evidence of positive selection because pairwise Ka/Ks estimates may be influenced by substitution saturation, alignment uncertainty, and lineage-specific evolutionary history. Nevertheless, the accelerated evolution of these genes suggests that they may represent genomic regions with reduced evolutionary constraints or enhanced sequence divergence. In particular, *ycf1* has been widely recognized as one of the most variable chloroplast genes among angiosperms and has been frequently applied in plant species discrimination and DNA barcode development [46, 47]. supporting its potential importance in *Hippeastrum* evolutionary studies.

The analysis of IR/SC boundary regions further demonstrated that chloroplast genome evolution in *Hippeastrum* was characterized by both structural conservation and lineage-specific modification. Most species exhibited highly similar IR boundary configurations, indicating strong constraints on IR stability during evolution. The maintenance of conserved boundary positions involving genes such as *rpl22*, *rps19*, *rpl2*, and *ycf1* suggests that extensive structural rearrangements rarely occurred during *Hippeastrum* diversification. However, *H. rutilum* represented an exception, showing pronounced IR expansion and SSC contraction compared with other species. The expansion of the IR region resulted in substantial differences in the positions of genes around the IR/SSC junctions, including ndh-related genes, which differed markedly from the conserved *ycf1/ndhF*-associated boundary pattern observed in other *Hippeastrum* species. Interestingly, *H. rutilum* also showed elevated evolutionary signals in *ycf2*, suggesting that structural variation and accelerated sequence evolution may have occurred concurrently in this lineage. Such lineage-specific IR expansion has been reported in several plant groups and is considered an important mechanism contributing to chloroplast genome size variation and structural diversification. Therefore, *H. rutilum* may represent a distinct evolutionary

trajectory within *Hippeastrum*, characterized by both genomic structural remodeling and sequence-level divergence.

Comparative genome alignment further revealed heterogeneous evolutionary patterns among *Hippeastrum* cpDNA. The mVISTA analysis showed high sequence conservation and strong collinearity, especially in protein-coding genes and rRNA regions, reflecting strong functional constraints during evolution. In contrast, non-coding regions, intergenic spacers, and several boundary-associated regions exhibited higher sequence divergence. Variable regions, including t*rnK-UUU/matK, rps16-trnQ-UUG, atpF-atpH, rpl32-trnL, ndhF*, and *ycf1*, may represent evolutionary hotspots within *Hippeastrum* cpDNA. These divergent regions could provide potential molecular markers for species identification and phylogenetic studies. The accelerated evolutionary pattern of *ycf1* inferred from Ka/Ks analysis was consistent with the sequence divergence detected by mVISTA analysis, indicating that *ycf1* represents a potential evolutionary hotspot in *Hippeastrum* cpDNA. The phylogenetic analysis based on 37 cpDNA further placed *Hippeastrum* within the broader evolutionary framework of Amaryllidaceae and revealed lineage differentiation within the genus. The three major subfamilial clades recovered in Amaryllidaceae were consistent with previous taxonomic classifications, supporting the reliability of cpDNA for resolving higher-level evolutionary relationships. Within *Hippeastrum*, three major clades were identified, showing partial correspondence with geographic distribution and floral morphological characteristics. Species distributed in the Andes and adjacent regions were clustered together, whereas species from the Atlantic Forest region formed another lineage. These geographic patterns suggest that historical biogeographic isolation and regional environmental conditions may have contributed to chloroplast lineage divergence. In addition, differences in floral morphology among clades, including tepal width, flower size, and growth habits, indicate that morphological diversification may have accompanied evolutionary differentiation within *Hippeastrum*. However, because cpDNA represent only the maternal evolutionary history and some nodes showed relatively limited support, these

relationships should be interpreted cautiously and further evaluated using nuclear genomic data and broader taxonomic sampling.

Overall, our results reveal that *Hippeastrum* chloroplast genome evolution is characterized by a combination of strong conservation and localized divergence. The conserved genome structure, gene content, and functional regions reflect long-term evolutionary constraints imposed by essential chloroplast functions, whereas variable regions, accelerated genes, and lineage-specific IR boundary changes provide evidence of ongoing evolutionary differentiation. The integration of selective pressure analysis, IR boundary comparison, genome-wide alignment, and phylogenetic reconstruction provides a comprehensive framework for understanding how cpDNA contribute to species diversification in *Hippeastrum*. These findings not only expand our understanding of chloroplast genome evolution in Amaryllidaceae but also provide valuable genetic resources for species identification, phylogenetic studies, and conservation of *Hippeastrum* germplasm.

Conclusion

In this study, chloroplast genomes of eight *Hippeastrum* species were sequenced, assembled, and annotated, revealing a conserved genomic framework and lineage-specific evolutionary patterns within the genus. *Hippeastrum* cpDNA showed high conservation in genome size, GC content, gene composition, and organization, reflecting strong functional constraints during evolution.

IR boundary analysis demonstrated that most species maintained stable SC/IR structures, whereas *H. rutilum* exhibited distinct IR expansion and SSC contraction, indicating lineage-specific structural remodeling. Selective pressure analysis revealed that most chloroplast genes were under strong purifying selection, while *ycf1* and *ycf2* showed relatively accelerated sequence divergence in specific lineages, representing potential evolutionary hotspots.

mVISTA analysis further revealed that conserved coding and rRNA regions maintain chloroplast genome stability, whereas non-coding regions, IR/SC boundary regions, and variable loci such as *ycf1* and *ndhF* contributed to species-level differentiation. Phylogenetic analysis based on 37 cpDNA resolved the evolutionary relationships within Amaryllidaceae and identified three major clades within *Hippeastrum*, with patterns partially associated with geographic distribution and floral traits.

Overall, this study demonstrates that *Hippeastrum* chloroplast genome evolution follows a pattern of high conservation accompanied by localized divergence, providing new insights into chloroplast genome evolution, species identification, and germplasm conservation in *Hippeastrum*.

Methods

Plant Materials, DNA Extraction, and Sequencing

The plant materials used in this study were obtained from the Danzhou Experimental Station of the Institute of Tropical Crop Varieties and Resources, Chinese Academy of Tropical Agricultural Sciences. Eight wild species of the genus *Hippeastrum* were selected (Figure 8). Fresh leaves were collected from each species, immediately frozen in liquid nitrogen, and stored at -80°C. Genomic DNA was extracted from the collected samples and used for sequencing. Whole-genome sequencing was performed on the DNBSEQ platform. After sequencing, the raw data were filtered to remove adapter sequences and low-quality reads. Complete cpDNA resequencing was performed on the DNBSEQ platform. DNA libraries with an average insert size of approximately 300-400 bp were constructed and sequenced in paired-end mode (PE150). Following sequencing, raw reads were subjected to quality control using SOAPnuke software. The filtering procedures included removal of adapter sequences, elimination of low-quality reads, and exclusion of reads containing excessive ambiguous bases (N) or polyX sequences. The specific parameters were set as follows: -n 0.01 -l 20 -

q 0.5 --adaMR 0.25 --polyX 50 --minReadLen 150. After filtering, high-quality clean reads were obtained and used for subsequent genome assembly analyses.

The raw sequencing data generated in this study have been deposited in the NCBI Sequence Read Archive (SRA) under BioProject accession PRJNA1476836.

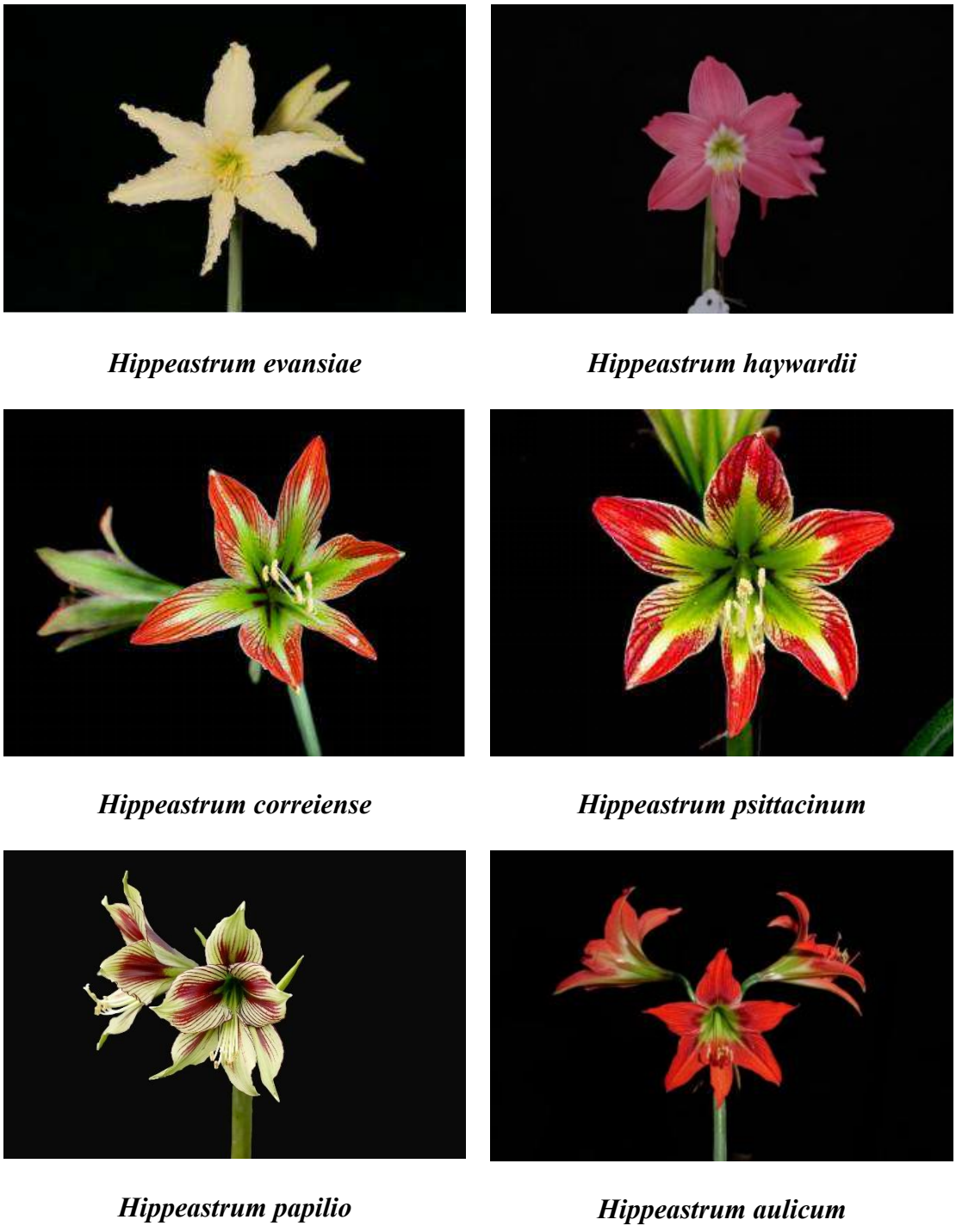


Figure 8 Wild species of the genus *Hippeastrum*

Chloroplast Genome Assembly and Annotation

Based on the obtained clean reads, cpDNA were assembled using GetOrganelle (v3.15.5). Paired-end sequencing data (PE150) were used as input, and the plant chloroplast genome mode provided by GetOrganelle (-F embplant_pt) was applied. Multiple k-mer sizes (-k 21,45,65,85,105) were specified to improve assembly completeness and accuracy, and the random seed parameter was set to 15 (-R 15). After assembly, the assembly graphs generated by GetOrganelle were visualized using Bandage v0.8.1 to evaluate the structural integrity of the cpDNA. Particular attention was paid to whether the assemblies formed a single closed circular structure. Node coverage depth information was examined to assess the copy number, connectivity, and coverage consistency of the IRs. Chloroplast genome sequencing depth was also calculated, revealing generally high coverage across all samples, with a minimum average depth of 98.3×, meeting the quality requirements for accurate chloroplast genome assembly. For assemblies containing gaps, a previously published *Hippeastrum* chloroplast genome (GenBank accession number: MT762362) was used as a seed sequence, and guided reassembly was performed in GetOrganelle by enabling the -s parameter to recover potentially missing regions. Ultimately, only cpDNA that were structurally complete, successfully circularized, and exhibited uniform coverage depth were retained for subsequent annotation and comparative analyses.

Based on the assembled cpDNA and a reference *Hippeastrum* chloroplast genome (GenBank accession number: MT762362), cpDNA annotation was conducted using the online platform CPGAVAS2. The cpDNA FASTA files assembled in this study, together with the annotated GenBank file of the reference species, were uploaded for template-based automatic annotation, which generated corresponding annotation results and tbl files. Annotation information, including gene names, genomic coordinates, and strand orientation, was subsequently manually curated. To further ensure annotation accuracy, the reference cpDNA sequences (*Hippeastrum vittatum*, MT762362; *H.*

reticulatum, MT701523) were individually aligned with the cpDNA generated in this study using MEGA software, and any potential annotation discrepancies were manually corrected.

Sequence Feature Analysis

Basic genome characteristics were calculated using SeqKit v2.8.0. Based on the annotation results, functional genes were classified and summarized. Protein-coding genes were grouped according to their functional categories and gene families. The number and types of genes containing introns, the locations of trans-spliced genes (such as *rps12*), and gene loss relative to the reference genome were also analyzed.

Codon usage bias analysis was conducted based on complete coding sequences (CDSs), and the RSCU was calculated for each sample. Repeat sequence analysis included the identification of simple sequence repeats (cpSSRs) and tandem repeats. cpSSRs were detected using MISA (https://webblast.ipk-gatersleben.de/misa/) with standard parameters. The minimum repeat numbers were set as follows: 10 for mononucleotide repeats, 6 for dinucleotide repeats, and 5 for tri-, tetra-, penta-, and hexanucleotide repeats. When the distance between two SSR loci was less than or equal to 100 bp, the loci were classified as compound SSRs. Based on these analyses, the types and distribution of SSRs, as well as changes in their abundance, were summarized.

Selection Pressure Analysis

Base substitutions that result in amino acid changes are defined as nonsynonymous mutations, whereas those that do not result in amino acid changes are defined as synonymous mutations. Nonsynonymous mutations are generally influenced by natural selection. The ratio of Ka to Ks was used to infer selective pressures acting on genes. A Ka/Ks ratio greater than 1 indicates that a gene may be subject to positive selection; a ratio equal to 1 suggests that nonsynonymous and synonymous substitutions occur at similar rates, implying neutral evolution; whereas a ratio less than 1 indicates that nonsynonymous substitutions occur at a lower rate than synonymous substitutions, suggesting purifying (negative) selection. To evaluate the selective pressure acting on chloroplast protein-coding

genes in *Hippeastrum*, Ka, Ks, and Ka/Ks ratios were calculated using the KaKs module implemented in CPStools. *Hippeastrum vittatum* was selected as the reference species, and 73 shared protein-coding genes among 12 *Hippeastrum* cpDNA were identified for analysis. The corresponding coding sequences were extracted and compared among species to estimate Ka, Ks, and Ka/Ks values at the codon level. The calculated Ka/Ks ratios were visualized using TBtools to evaluate differences in evolutionary rates and selective constraints among chloroplast genes.

Comparative Genome Analysis

The sizes and organizational structures of the cpDNA from twelve *Hippeastrum* species were compared, and differences in the IR boundaries were analyzed. Contraction and expansion of the IR regions were examined using the online tool IRscope (*https:*//irscope.shinyapps.io/irapp/). Gene annotation files (GenBank format) of the twelve *Hippeastrum* cpDNA were uploaded to IRscope for analysis, and visualizations were generated to compare the IR/LSC/SSC boundary structures among the different species.

Comparative analysis of chloroplast genome sequences was performed using the mVISTA online platform (*https:*//genome.lbl.gov/vista/index.shtml). The complete chloroplast genome sequence of *Hippeastrum evansiae* was selected as the reference genome, and the cpDNA of 11 additional *Hippeastrum* species were compared to investigate sequence conservation and divergence within the genus. In addition, ten cpDNA from other genera of Amaryllidaceae were selected for intergeneric comparison.

To evaluate the distribution pattern of sequence variation across the cpDNA of *Hippeastrum*, nucleotide diversity (Pi) analysis was performed based on complete chloroplast genome sequences. The chloroplast genome sequences of 12 *Hippeastrum* species were first standardized for orientation using the Seq module implemented in CPStools (*https://*github.com/Xwb7533/CPStools) to ensure consistent directionality among sequences. Because the chloroplast genome of *Hippeastrum rutilum* (MT937175) exhibited lineage-specific structural variation, including gene loss and rearrangement

within the SSC region, it was excluded from the whole-genome sliding-window analysis to avoid potential bias caused by large-scale structural divergence. The remaining 11 cpDNA were aligned using MAFFT v7.310 with the --auto strategy. The aligned sequences were subsequently imported into DnaSP v6 (Rozas et al., 2017) for sliding-window nucleotide diversity analysis. A window size of 600 bp and a step size of 200 bp were applied to estimate nucleotide variation across the cpDNA. The resulting Pi values were visualized using Python scripts to identify highly variable regions within the *Hippeastrum* cpDNA. To further evaluate nucleotide diversity at the gene level, shared protein-coding sequences (CDSs) among the 12 *Hippeastrum* cpDNA were extracted from annotated GenBank files using the Pi module of CPStools. Each homologous CDS region was aligned using MAFFT v7.310 with default parameters, and nucleotide diversity values for individual genes were calculated using CPStools. The obtained gene-level Pi values were visualized using R scripts to identify rapidly evolving genes and potential molecular marker regions.

Phylogenetic Analysis

To investigate the phylogenetic position of *Hippeastrum* within Amaryllidaceae, a broad taxon sampling strategy was employed. The dataset included eight wild *Hippeastrum* cpDNA generated in this study, four previously published *Hippeastrum* genomes retrieved from the NCBI GenBank database (*Hippeastrum alberti* MT701522, *Hippeastrum vittatum* MT762362, *Hippeastrum reticulatum* MT701523, and *Hippeastrum rutilum* MT937175), and representative cpDNA from related taxa within Amaryllidoideae, Allioideae, and Agapanthoideae. *Acorus tatarinowii* (NC045294) was used as the outgroup to root the phylogenetic tree. This resulted in a total dataset of 37 cpDNA, representing multiple taxonomic levels within Amaryllidaceae, enabling a comprehensive phylogenetic reconstruction across family-level lineages.

Multiple sequence alignment was performed using MAFFT v7.505 implemented in PhyloSuite v2.0. The alignment was conducted using the "--auto" strategy, which automatically selected the L-

INS-i algorithm under default parameters. Sequences were aligned in input order (--inputorder), and no manual editing was performed. Gap characters and missing data were retained in the final alignment, and no ambiguous or poorly aligned regions were removed prior to phylogenetic reconstruction. The complete alignment was directly used for downstream analysis.

Phylogenetic trees were constructed using the PhyloSuite platform. Maximum likelihood (ML) analysis was performed with IQ-TREE v3.0.1. The nucleotide substitution model was selected using the ModelFinder module, and the best-fit model was chosen according to the Bayesian Information Criterion (BIC), with TVM+F+G4 identified as the optimal model. Branch support was assessed using the Ultrafast Bootstrap method with 5000 replicates. *Acorus tatarinowii* was specified as the outgroup to root the tree. The resulting tree file was visualized in FigTree v1.4.4 and further edited and refined using iTOL (https://itol.embl.de/).

Abbreviations

cpDNA: Chloroplast genomes; ITS: Internal transcribed spacer; ILS: incomplete lineage sorting; SSR: Simple sequence repeat; IRs: Inverted repeats; LSC: Large single-copy; SSC: Small single-copy; RSCU: relative synonymous codon usage; Ka: nonsynonymous substitution rate; Ks: synonymous substitution rate; ML: Maximum-likelihood; BIC: Bayesian Information Criterion.

Supplementary Material

Table S1 Statistics of clean data

Table S2 Codon usage in the chloroplast genomes of eight *Hippeastrum* species

Table S3 Simple sequence repeats (SSRs) identified in the chloroplast genome of *Hippeastrum* species

Table S4. Pairwise KaKs values of 73 chloroplast CDS genes between *Hippeastrum vittatum* and 11 other *Hippeastrum* species.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The datasets presented in this study can be found in online repositories. The raw sequencing reads have been deposited in the NCBI Sequence Read Archive (SRA) (https://www.ncbi.nlm.nih.gov/sra/) under BioProject accession number PRJNA1476836. The assembled and annotated chloroplast genome sequences have been deposited in GenBank, with accession numbers listed in the manuscript.

Funding

The author(s) declare financial support was received for the research and/or publication of this article. This research was financially supported by the Fundamental Research Funds for the Chinese Academy of Tropical Agricultural Sciences (Institute of Tropical Crop Genetic Resources) under the project "Collection, Conservation, and Cataloguing of Hippeastrum Germplasm Resources" (Project No. 1630032026030).

Authors' contributions

L.Y., Y.G. and X.L. designed the research. L.Y. and X.L. performed the experiments, as sembled the genomes, and analyzed the data. Y.G. and S.Z. identified the plant materials. Y.S and S.Z contributed to the methodology section. L.Y. drafted the manuscript. L.Y., X.L. and Y.G. revised the manuscript. Y.G supervised the project and secured funding. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

Reference

1. Wang Y, Chen D, He X, Shen J, Xiong M, Wang X, et al. Revealing the complex genetic structure of cultivated amaryllis (Hippeastrum hybridum) using transcriptome-derived microsatellite markers. Scientific Reports. 2018;8(1):10645.

2. Marasek-Ciolakowska A, Sochacki D, Marciniak P. Breeding aspects of selected ornamental bulbous crops. Agronomy. 2021;11(9):1709.

3. Meerow A, Broschat T, Kane M. Breeding of new Hippeastrum cultivars using diploid species. Acta horticulturae. 1991;325:583-9.

4. Meerow A. The florida series of Hybrid Amaryllis: Five new Hippeastrum cultivars. HortScience. 2014;49:1102-7.

5. Kamenetsky R, editor Production of flower bulbs in regions with warm climates2005: International Society for Horticultural Science (ISHS), Leuven, Belgium; 2005.

6. Shao L, Yang L, Li X, Zhou L, Zhu J, Zhang Y. From bulb development to postharvest treatments: advances in Hippeastrum spp. research and industry applications. Ornamental Plant Research. 2025;5(1):e026.

7. Sebben C, Giordani RB, de Andrade JP, Berkov S, Osorio EJ, Sobral M, et al. New lycosinine derivative from Hippeastrum breviflorum. Revista Brasileira de Farmacognosia. 2015;25(4):353-5.

8. García N, Meerow AW, Arroyo-Leuenberger S, Oliveira RS, Dutilh Julie H, Soltis Pamela S, et al. Generic classification of Amaryllidaceae tribe Hippeastreae. TAXON. 2019;68(3):481-98.

9. Cândido RS, Lopes RC, Mendonça CBF, Gonçalves-Esteves V. Pollen morphology of Hippeastrum (Amaryllidaceae): implications for taxonomic delimitation and systematics. Brazilian Journal of Botany. 2025;48(1):63.

10. Silvia A. The Chromosomes of Hippeastrum, Amaryllis and Phycella (Amaryllidaceae). Kew Bulletin. 1982;37(2):211-6.

11. Naranjo CA, Poggio L. A Comparison of Karyotype, Ag-NOR Bands and DNA Content in Amaryllis and Hippeastrum (Amaryllidaceae). Kew Bulletin. 1988;43(2):317-25.

12. Poggio L, González GE, Naranjo CA. Chromosome studies in Hippeastrum (Amaryllidaceae): variation in genome size. Botanical Journal of the Linnean Society. 2007;155(2):171-8.

13. Poggio L, Realini MF, Fourastié MF, García AM, González GE. Genome downsizing and karyotype constancy in diploid and polyploid congeners: a model of genome size variation. AoB PLANTS. 2014;6.

14. Williams M, Dudley TR. CHROMOSOME COUNT FOR HIPPEASTRUM IGUAZUANUM. TAXON. 1984;33(2):271-5.

15. Garcia N, Folk RA, Meerow AW, Chamala S, Gitzendanner MA, Oliveira RS, et al. Deep reticulation and incomplete lineage sorting obscure the diploid phylogeny of rain-lilies and allies (Amaryllidaceae tribe Hippeastreae). Mol Phylogenet Evol. 2017;111:231-47.

16. Li DZ, Gao LM, Li HT, Wang H, Ge XJ, Liu JQ, et al. Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. Proc Natl Acad Sci U S A. 2011;108(49):19641-6.

17. Clement WL, Donoghue MJ. Barcoding success as a function of phylogenetic relatedness in Viburnum, a clade of woody angiosperms. BMC Evolutionary Biology. 2012;12(1):73.

18. Besse P, Da Silva D, Grisoni M. Plant DNA Barcoding Principles and Limits: A Case Study in the Genus Vanilla. Methods Mol Biol. 2021;2222:131-48.

19. Chen Z, Gao L, Wang H, Feng S. Molecular Identification and Phylogenetic Analysis of Cymbidium Species (Orchidaceae) Based on the Potential DNA Barcodes matK, rbcL, psbA-trnH, and Internal Transcribed Spacer. Agronomy. 2024;14(5):933.

20. Ravi V, Khurana JP, Tyagi AK, Khurana P. An update on chloroplast genomes. Plant Systematics and Evolution. 2008;271(1):101-22.

21. Daniell H, Lin CS, Yu M, Chang WJ. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biol. 2016;17(1):134.

22. Ohyama K, Fukuzawa H, Kohchi T, Shirai H, Sano T, Sano S, et al. Chloroplast gene organization deduced from complete sequence of liverwort Marchantia polymorpha chloroplast DNA. Nature. 1986;322(6079):572-4.

23. Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, et al. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. Embo j. 1986;5(9):2043-9.

24. Lu R-S, Li P, Qiu Y-X. The Complete Chloroplast Genomes of Three Cardiocrinum (Liliaceae) Species: Comparative Genomic and Phylogenetic Analyses. Frontiers in Plant Science. 2017;Volume 7 - 2016.

25. Loeuille B, Thode V, Siniscalchi C, Andrade S, Rossi M, Pirani JR. Extremely low nucleotide diversity among thirty-six new chloroplast genome sequences from Aldama (Heliantheae, Asteraceae) and comparative chloroplast genomics analyses with closely related genera. PeerJ. 2021;9:e10886.

26. Li Y, Zhang L, Wang T, Zhang C, Wang R, Zhang D, et al. The complete chloroplast genome sequences of three lilies: genome structure, comparative genomic and phylogenetic analyses. J Plant Res. 2022;135(6):723-37.

27. Wang X, Bai S, Zhang Z, Zheng F, Song L, Wen L, et al. Comparative analysis of chloroplast genomes of 29 tomato germplasms: genome structures, phylogenetic relationships, and adaptive evolution. Frontiers in Plant Science. 2023;Volume 14 - 2023.

28. Parks M, Cronn R, Liston A. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biol. 2009;7:84.

29. 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.

30. Rather SA, Wang K, Wang T, Liu H, Schneider H, fls. Comparative chloroplast genome analysis reveals powerful barcodes for combatting illegal logging of CITES-listed threatened Asian rosewoods (Dalbergia, Leguminosae, Papilionoideae). Botanical Journal of the Linnean Society. 2024;208(4):347-68.

31. Hori TA, Hayashi A, Sasanuma T, Kurita S. Genetic variations in the chloroplast genome and phylogenetic clustering of Lycoris species. Genes Genet Syst. 2006;81(4):243-53.

32. Contreras-Diaz R, Arias-Aburto M, van den Brink L. Characterization of the complete chloroplast genome of Zephyranthes phycelloides (Amaryllidaceae, tribe Hippeastreae) from Atacama region of Chile. Saudi J Biol Sci. 2022;29(1):650-9.

33. He L, Zou B, Gao H, Liu L, Wu Y. The complete chloroplast genome of Hippeastrum rutilum (Ker-Gawl.) Herb. Mitochondrial DNA B Resour. 2021;6(3):1141-3.

34. Zhang F, Wang N, Cheng G, Shu X, Wang T, Zhuang W, et al. Comparative Chloroplast Genomes of Four Lycoris Species (Amaryllidaceae) Provides New Insight into Interspecific Relationship and Phylogeny. Biology (Basel). 2021;10(8).

35. Könyves K, Bilsborrow J, David J, Culham A. The complete chloroplast genome of Narcissus poeticus L. (Amaryllidaceae: Amaryllidoideae). Mitochondrial DNA B Resour. 2018;3(2):1137-8.

36. Namgung J, Do HDK, Kim C, Choi HJ, Kim JH. Complete chloroplast genomes shed light on phylogenetic relationships, divergence time, and biogeography of Allioideae (Amaryllidaceae). Sci Rep. 2021;11(1):3262.

37. Li P, Ren M, Zhu Q, Zhang Y, Xu H, Wang Z, et al. The complete chloroplast genome of Hippeastrum vittatum (Amaryllidaceae). Mitochondrial DNA B Resour. 2020;5(3):3521-3.

38. Liu XF, Sun YB, Zhu GF, Huang LL, Yu B. Complete chloroplast genomes and comparative analyses of Hippeastrum 'milady', Hippeastrum albertii and Hippeastrum reticulatum (Amaryllidaceae). PLoS One. 2022;17(8):e0271335.

39. Huang B. The complete chloroplast genome sequence of Hippeastrum rutilum (Amaryllidoideae). Mitochondrial DNA B Resour. 2020;5(3):3387-8.

40. Huang H, Shi C, Liu Y, Mao SY, Gao LZ. Thirteen Camellia chloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships. BMC Evol Biol. 2014;14:151.

41. Kaila T, Chaduvla PK, Rawal HC, Saxena S, Tyagi A, Mithra SVA, et al. Chloroplast Genome Sequence of Clusterbean (Cyamopsis tetragonoloba L.): Genome Structure and Comparative Analysis. Genes (Basel). 2017;8(9).

42. Zhang X, Rong C, Qin L, Mo C, Fan L, Yan J, et al. Complete Chloroplast Genome Sequence of Malus hupehensis: Genome Structure, Comparative Analysis, and Phylogenetic Relationships. Molecules. 2018;23(11).

43. Li W, Zhang C, Guo X, Liu Q, Wang K. Complete chloroplast genome of Camellia japonica genome structures, comparative and phylogenetic analysis. PLoS One. 2019;14(5):e0216645.

44. Garrick RC, Bonatelli IAS, Hyseni C, Morales A, Pelletier TA, Perez MF, et al. The evolution of phylogeographic data sets. Molecular Ecology. 2015;24(6):1164-71.

45. Hu Ying WX, Zhang Xinxin, Zhou Wei, Chen Xiaoyang, Hu Xinsheng. Advancing phylogeography with chloroplast DNA markers. Biodiv Sci. 2019;27(2):219-34.

46. Liu HY, Yu Y, Deng YQ, Li J, Huang ZX, Zhou SD. The Chloroplast Genome of Lilium henrici: Genome Structure and Comparative Analysis. Molecules. 2018;23(6).

47. Li B, Lin F, Huang P, Guo W, Zheng Y. Development of nuclear SSR and chloroplast genome markers in diverse Liriodendron chinense germplasm based on low-coverage whole genome sequencing. Biological Research. 2020;53(1):21.

768

Supplementary Files

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

- TableS1Statisticsofcleandata.xlsx
- TableS2CodonusageinthechloroplastgenomesofeightHippeastrumspecies.xlsx
- TableS4PairwiseKaKsvaluesof73chloroplastCDSgenesbetweenHippeastrumvittatumand11otherHippeastrumspecies..xlsx
- TableS3SimplesequencerepeatsSSRsidentifiedinthechloroplastgenomeofHippeastrumspecies.xlsx