

Complete Chloroplast Genome of a Montane Plant *Spathoglottis aurea* Lindl.: Comparative Analyses and Phylogenetic Relationships among members of Tribe Collabieae

Nurul Shakina Mohd Talkah¹, Jasim Haider Mahmud Jasim¹, Farah Alia Nordin¹, Ahmad Sofiman Othman^{1,2*}

1 School of Biological Sciences, Universiti Sains Malaysia, 11800 Gelugor, Penang, Malaysia

2 Center of Chemical Biology, Universiti Sains Malaysia, 11900 Bayan Lepas, Penang, Malaysia

* sofiman@usm.my

Abstract

The yellow-flowered *Spathoglottis aurea* (tribe Collabieae; family Orchidaceae) is native to the mountainous areas of Peninsular Malaysia. The species is well known as an ornamental plant and for its role in artificial hybrid breeding. There is an interesting evolutionary relationship between *S. aurea* and the geographically isolated *S. microchilina* from Borneo that has encouraged further study of the *S. aurea* populations, but the genomic resource for *S. aurea* has not yet been reported. The present study reports the first work to characterize a chloroplast genome among the *Spathoglottis* genus. The complete chloroplast (cp) genome of *S. aurea* was assembled from a sequence generated by the Illumina platform and analysed in comparison with other Collabieae species available in the GenBank database. The cp genome of *S. aurea* is 157,957 base pairs (bp) in length with guanine-cytosine (GC) content of 37.3%. The genome possessed a typical quadripartite chloroplast genome structure with large single-copy (LSC) (86,888 bp), small single-copy (SSC) (18,125 bp) and inverted repeat (IR) (26,472 bp) sequences. A total of 134 genes were annotated, with 88 protein coding genes (PCGs), 38 transfer RNA (tRNA) genes and eight ribosomal RNA (rRNA) genes. Overall, 80 simple sequence repeats (SSR) or microsatellites were identified. Comparative analysis with other Collabieae species revealed high conservation in the cp genome arrangements with minimal difference in genome lengths. However, several mutational hotspots were also detected, with high potential to be developed as genetic markers for phylogenetic analysis. Characterization of the *S. aurea* cp genome revealed its conserved nature without gene loss or rearrangements when compared to other species of the Collabieae tribe. Phylogenetic analysis of Collabieae species also revealed that *S. aurea* has a distant evolutionary relationship to other members of the Collabieae species, despite the presence of problematic genera such as *Phaius* and *Cephalantheropsis*.

Introduction

The chloroplast genome is about 120-200 kilobases (kb) in size. In term of gene size, gene content, gene arrangements and genomic structure, it has frequently been shown to be conserved, except for a number of taxa (1,2). These characteristics, together with the absence of recombination and inheritance through a single lineage, make it a suitable marker for phylogeny and genetic variability studies of terrestrial plants. It has been shown to be a

powerful molecular tool for taxonomy study (3), phylogenetics (4,5) and conservation genetics.

Collabieae is a tribe of orchids within the large family of Orchidaceae. It was first described by Pfitzer in 1887 (6), based on the genus *Collabium* and predominantly consists of terrestrial orchids. The high diversity of morphological characteristics such as vegetative and floral variations within Collabieae has made it difficult for systematists to infer their evolutionary relationships. Despite this, based on a revision by Chase (7), among the Collabieae tribe the chloroplast genome has been reported for several taxa, such as *Calanthe* (8–11) and *Tainia* (12–14). Although the group has approximately 450–500 known species, scarce genomic data are available and there are few reports of taxonomic conflicts (9,15,16).

The genus *Spathoglottis* Blume of Collabieae tribe is a favourite of horticulture enthusiasts, because it has bright, colourful flowers in deep crimson, pink and golden yellow that bloom all year round, is easy to cultivate and quick to grow. The Plants of the World Online (POWO) database, facilitated by Kew Gardens, contains 49 accepted species of the *Spathoglottis* genus. The natural distribution of this genus has been recorded in the Pacific Islands and in tropical and subtropical Asia. Across the Malesian region there are 44 species that are widely distributed, with particularly dense gatherings in New Guinea (15). There are four *Spathoglottis* species recorded in Peninsular Malaysia with one infraspecific taxon: *S. aurea* Lindl., *S. plicata* Blume, *S. hardigiana* C.S.P. Parish & Rchb.f., *S. affinis* de Vriese and *S. plicata* var. *alba*. The first of these is a common mountain species native to Peninsular Malaysia, whose population is restricted to the top of mountains, where it usually grows on the granitic terrain of moist forests. It was first described by Lindley in 1850, based on a specimen from Mount Ophir (Gunung Ledang) in Johor (17) and was named *S. wrayi* 40 years later by Hooker, before receiving its present-day name of *S. aurea*. As a part of the *Spathoglottis* genus with a yellow flower complex, *S. aurea* is unique because of the influence of anthocyanin in its floral parts and leaves. Recent molecular (15) and morphological (18) studies have shown that *S. aurea* could be the closest ancestor of *S. microchilina* Kraenzl., which has only been reported in the Borneo region. Further population studies may confirm this relationship between *S. aurea* and *S. microchilina*.

The present study aimed to: (a) characterize the cp genome of *S. aurea*; (b) compare the *S. aurea* chloroplast genome structure with that of fellow species of the Collabieae tribe; (c) determine the phylogenetic positions of *S. aurea* within the Collabieae tribe in terms of all chloroplast protein coding genes. The availability of the cp genome in our dataset is expected to support further study on *S. aurea* populations and the phylogeny of the *Spathoglottis* yellow flower complex.

Materials and methods

Sampling, DNA Extraction, and Sequencing

Individual *S. aurea* plants were collected from Mount Jerai, Kedah, Malaysia and replanted in a greenhouse at Universiti Sains Malaysia. The voucher specimen was deposited in the Herbarium, School of Biological Sciences, Universiti Sains Malaysia (USMP11942). Fresh leaves from the cultivar were cleaned with running tap water and 70% ethanol and put in silica gel to dry. Genomic DNA was extracted from the dried leaves using 2X

cetyltrimethylammonium bromide (CTAB) according to the conventional method (19), with three modifications: (1) At the beginning of the protocol, after grinding the dry leaf tissue with liquid nitrogen, 750 µl CTAB-free buffer [200 mM Tris-HCl pH 8.0, EDTA 50 mM and NaCl 250 mM] and 5 µl β-mercaptoethanol were added. The samples were kept at -20°C for 10 minutes, before being subject to a centrifuge at maximum speed for 5 minutes and the supernatant discarded. These steps were repeated two or three times until a colourless or slightly yellow supernatant was obtained; (2) The chloroform – isoamyl alcohol (CIA) mixture (24:1) was added three times rather than twice; (3) After adding ice-cold isopropanol, the mixture was stored at -20°C for 1-2 hours. The samples were then centrifuged at 12,000 rpm for 10 minutes, the supernatant discarded, and the DNA pellet deposited carefully. A 600 µL wash buffer [2.5 M sodium acetate (NaAc); 76% ethanol, pH 5.0] was added to the DNA pellet to be washed and left at room temperature for 30 minutes, then centrifuged at 8,000 rpm for 1 minute and the supernatant discarded. The wash step was repeated twice, then the DNA pellet was allowed to dry completely at room temperature for 5 hours. Finally, the dried DNA was resuspended with 80-100µL of TE Buffer and incubated at 55°C for 30 minutes in a water bath. The genomic DNA was sequenced for 150bp paired end reads via the Illumina Novaseq 6000 using the TruSeq Nano DNA kit according to manufacturer protocol.

Chloroplast Genome Assembly, Assembly Validations, and Gene Annotation

De novo assembly of *S. aurea* chloroplast genome was done using GetOrganelle version 1.7.5. (20). The assembled chloroplast genome was validated by using coverage depth analysis. This analysis was done by mapping back the assembled genome to raw reads using ‘mem’ function of Burrows-Wheel aligner (21) and sorted using SamTools (22). Coverage depth for each nucleotide position in the assembled genome were retrieved using SamTools. The finalized chloroplast genome was then annotated using GeSeq (23) with Chloe function invoked. The open reading frame positions were confirmed by comparing them with reference genomes of closely related species using Genbank Blast function (24). The annotation table generated by GB2Sequin was edited according to the correct reading frame before being deposited in the Genbank. Visualization of *S. aurea* chloroplast genome was done from Genbank flatfile using CPGView platform (25).

Sequence Repeat Analysis and Codon Usage

Simple sequence repeats (SSRs) of the *S. aurea* cp genome were identified using MISA (26) with the parameter of 10 repeat units for mononucleotides, five repeat units for dinucleotides, four repeat units for trinucleotides and three repeat units for tetranucleotides, pentanucleotides and hexanucleotides. REPuter (27) was then used to detect forward, reverse, complement and palindrome repeats in the cp genome. Protein coding gene (PCG) sequences were extracted using Python scripts from the GenBank file. The PCG sequences were analysed for relative synonymous codon usage (RSCU) using the Bioinformatics web server (https://www.bioinformatics.org/sms2/codon_usage.html) (28) developed by Stothard (2000) and MEGA-X (29).

Comparative Analysis of Chloroplast Genome

The gene content and rearrangement in the *S. aurea* cp genome were compared with eight other Collabieae species available in the GenBank using progressiveMauve in Mauve version 2.4.0 (30). The genes available near inverted repeat boundaries were visualized using IRscope (31). Whole chloroplast genomes of eight Collabieae species were aligned using

MAFFT v7.511 (32). The sequence matrix was utilized to evaluate nucleotide diversity (π) using DNaSP6 (33) with sliding window analysis of 600bp window length and 200bp step size. The nucleotide diversity threshold was calculated by the sum of the average and double the variance of nucleotide diversity (π) (34).

Phylogenetic Analysis

Phylogenetic analysis was performed using 71 PCGs extracted from 29 different Collabieae species and one outgroup species (*Dendrobium sinense*) available in the GenBank. All PCGs were concatenated using MEGA-X and aligned by MAFFT v7.511. The best-fit model for sequence evolution was assessed using ModelFinder (35). Finally, a maximum likelihood (ML) tree was built based on the TVM+F+R2 model using IQ-Tree version 1.6.12 (36), operated on the Linux platform with 1000 bootstrap replicates.

Results

Characterization of the Chloroplast Genome of *Spathoglottis aurea*

Paired-end raw read sequencing of *S. aurea* produced 95,319,748 reads (7GB) with a GC content of 37.09% and a Phred score (Q30) of 86.48%. The raw reads were deposited in the Sequence Read Archive (SRA) of GenBank (accession number SRX1938371). A total of 652,534 reads were assembled to complete the chloroplast genome of 157,957 bp in length (GC content: 37.3%; GenBank accession number: OQ411079). The resulted assembly shows an average of 1,600X genome coverage (Figure S1) with a typical quadripartite structure of large single copy (LSC), small single copy (SSC) and a pair of inverted repeats (IR). Overall, 134 genes with 88 protein coding genes (PCG), eight ribosomal RNA (rRNA) genes and 38 transfer RNA (tRNA) genes were annotated in. There are 18 genes with intron with two of the contain two introns respectively. The two genes contained two introns are *psaI* and *clpP1*. Details on other genes present in *S. aurea* cp genome and their functionalities are listed in Table 1. The visualization of circularized chloroplast genome is shown in Figure 1.

The *S. aurea* cp genome contains 13 protein coding cis-splicing genes, of which two have two introns (*psaI*; also named *ycf3* and *clpP1*). The trans-splicing genes, *rps12*, consist of exons from three different regions; the 5'-end exon was found in the LSC region and the intron, and the 3'-end exon was located in the IR region. Figure 2 shows the list of cis-splicing and trans-splicing genes with their positions in the genome. There are 20,690 codons in the cp genome (Table S1). Codon usage analysis from *S. aurea* PCGs revealed codons ending with A/T nucleotides generally have higher usage compared to those ending with G/C nucleotides (Figure 3). The amino acid that was most coded in the genome is Leucine, with 2,118 codons. The codon occurring the most is GAA which is 871 times.

Table 1. Genes present in the *S. aurea* chloroplast genome. Genes with bold letters: contain intron, genes with bold and underlined letters: contain two introns.

Category for genes	Group of genes	Names of genes
Transcription and translation-related genes	Transfer RNAs	<i>trnA-UGC</i> (x2), <i>trnC-GCA</i> , <i>trnF-GAA</i> , <i>trnG-GCC</i> , <i>trnG-UCC</i> , <i>trnH-GUG</i> , <i>trnI-CAU</i> (x2), <i>trnI-GAU</i> (x2), <i>trnL-CAA</i> (x2), <i>trnL-UAA</i> , <i>trnL-UAG</i> , <i>trnM-CAU</i> , <i>trnN-GUU</i> (x2), <i>trnR-ACG</i> (x2), <i>trnR-UCU</i> , <i>trnS-GGA</i> , <i>trnT-GGU</i> , <i>trnV-GAC</i> (x2), <i>trnD-GUC</i> , <i>trnE-UUC</i> , <i>trnK-UUU</i> , <i>trnN-GUU</i> , <i>trnP-UGG</i> , <i>trnQ-UUG</i> , <i>trnS-GCU</i> , <i>trnS-UGA</i> , <i>trnT-UGU</i> , <i>trnV-UAC</i> , <i>trnW-CCA</i> , <i>trnM-CAU</i> , <i>trnY-GUA</i> ,
	Ribosomal RNAs	<i>rrn16</i> (x2), <i>rrn23</i> (x2), <i>rrn4.5</i> (x2), <i>rrn5</i> (x2)
	Ribosomal proteins (small subunit)	<i>rps11</i> , <i>rps12</i> (x2), <i>rps14</i> , <i>rps15</i> , <i>rps16</i> , <i>rps18</i> , <i>rps19</i> (x2), <i>rps2</i> , <i>rps3</i> , <i>rps4</i> , <i>rps7</i> (x2), <i>rps8</i>
	Ribosomal proteins (large subunit)	<i>rpl14</i> , <i>rpl16</i> , <i>rpl2</i> (x2), <i>rpl20</i> , <i>rpl22</i> , <i>rpl23</i> (x2), <i>rpl32</i> , <i>rpl33</i> , <i>rpl36</i>
	Translational initiation factor	<i>infA</i>
	DNA dependent RNA polymerase	<i>rpoA</i> , <i>rpoB</i> , <i>rpoC1</i> , <i>rpoC2</i>
Photosynthesis-related genes	NADH dehydrogenase	<i>ndhA</i> , <i>ndhB</i> (x2), <i>ndhC</i> , <i>ndhD</i> , <i>ndhE</i> , <i>ndhF</i> , <i>ndhG</i> , <i>ndhH</i> , <i>ndhI</i> , <i>ndhJ</i> , <i>ndhK</i>
	Photosystem I	<i>psaA</i> , <i>psaB</i> , <i>psaC</i> , <i>psaI</i> , <i>psaJ</i>
	Photosystem II	<i>psbA</i> , <i>psbB</i> , <i>psbC</i> , <i>psbD</i> , <i>psbE</i> , <i>psbF</i> , <i>psbH</i> , <i>psbI</i> , <i>psbJ</i> , <i>psbK</i> , <i>psbL</i> , <i>psbM</i> , <i>psbT</i> , <i>psbZ</i> , <i>pbf1</i>
	Cytochrome b/f complex	<i>petA</i> , <i>petB</i> , <i>petD</i> , <i>petG</i> , <i>petL</i> , <i>petN</i>
	RubisCO	<i>rbcL</i>
	ATP synthase	<i>atpA</i> , <i>atpB</i> , <i>atpE</i> , <i>atpF</i> , <i>atpH</i> , <i>atpI</i>
Unknown genes	Hypothetical chloroplast reading frame	<i>ycf1</i> (x2), <i>ycf2</i> (x2), <i>pafl</i> , <i>paflI</i> ,
Other genes	Maturase	<i>matK</i>
	Protease	<i>clpP1</i>
	Envelope membrane protein	<i>cemA</i>
	Subunit of Acetyl-CoA carboxylase	<i>accD</i>
	C-type cytochrome synthesis	<i>cemA</i>

Spathoglottis aurea

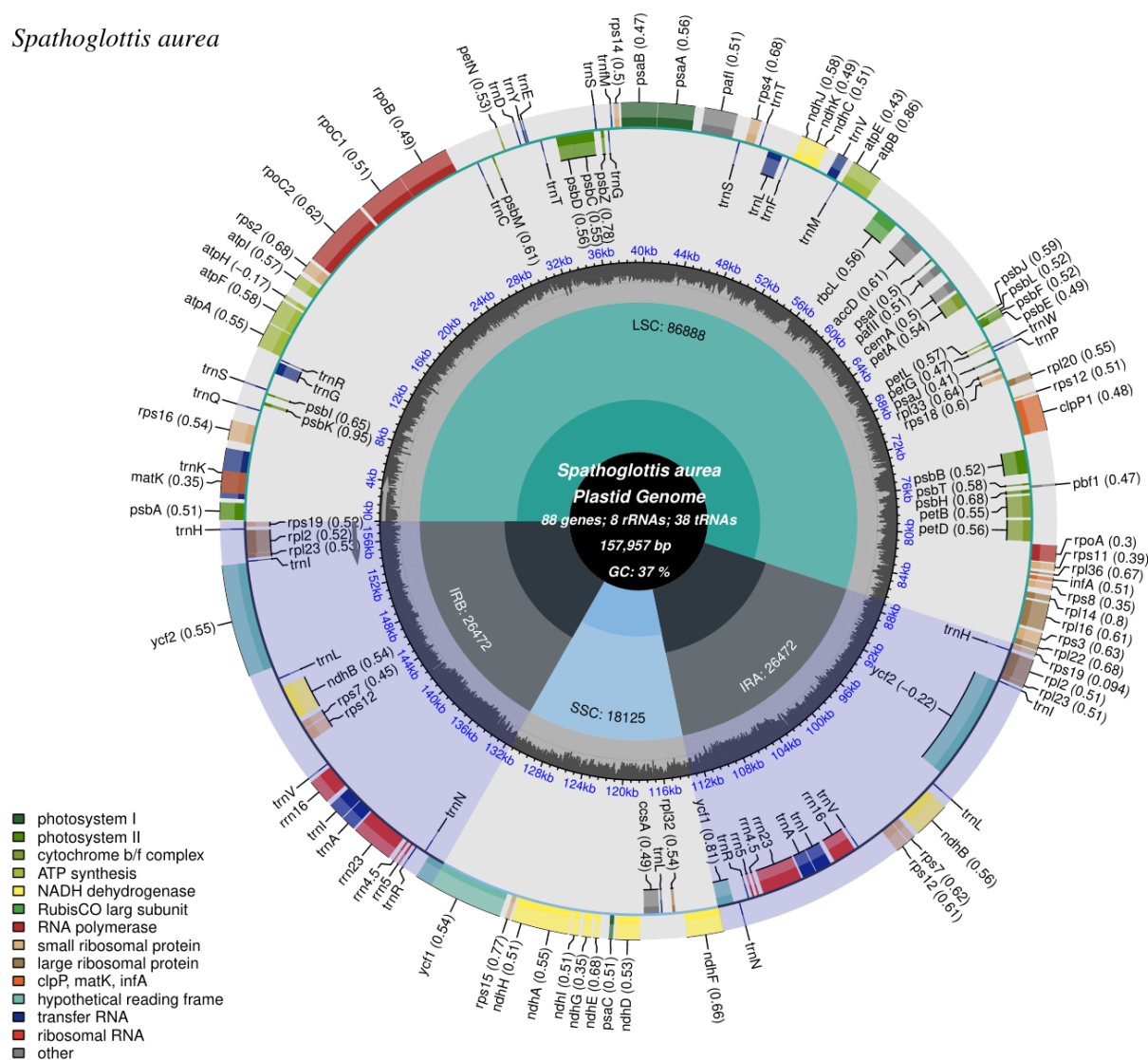


Fig. 1. The figure shows visualization of *S. aurea* chloroplast genome (OQ411079) generated by CPGView. The small single-copy (SSC), inverted repeat (IRa and IRb), and large single-copy (LSC) regions were labelled respectively. The genes shown on the most outer track were color coded according to their functional groups with legends available on the left bottom. The genes that transcribed in clockwise direction positioned in inner lines while the genes that transcribed in anticlockwise directions are positioned in outer lines. The codon usage bias is displayed in the parenthesis after the gene name.

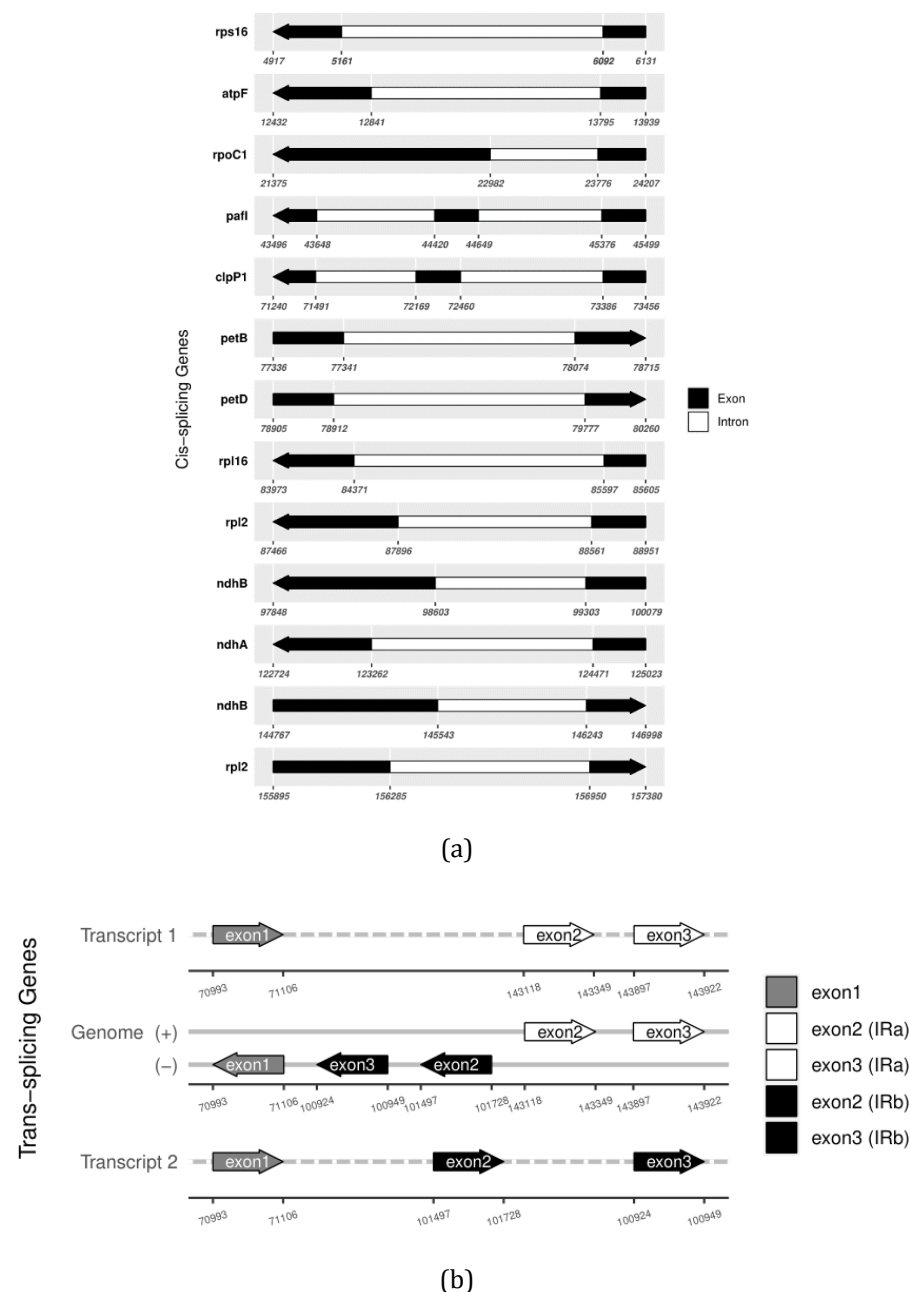


Fig. 2. The figure shows genes splicing positions in *S. aurea* chloroplast genome. (a) The genes with cis-splicing regions were arranged according to their positions in chloroplast genome. (b) Trans-splicing gene *rps12* with three unique exons. The gene is duplicated in inverted repeat regions.

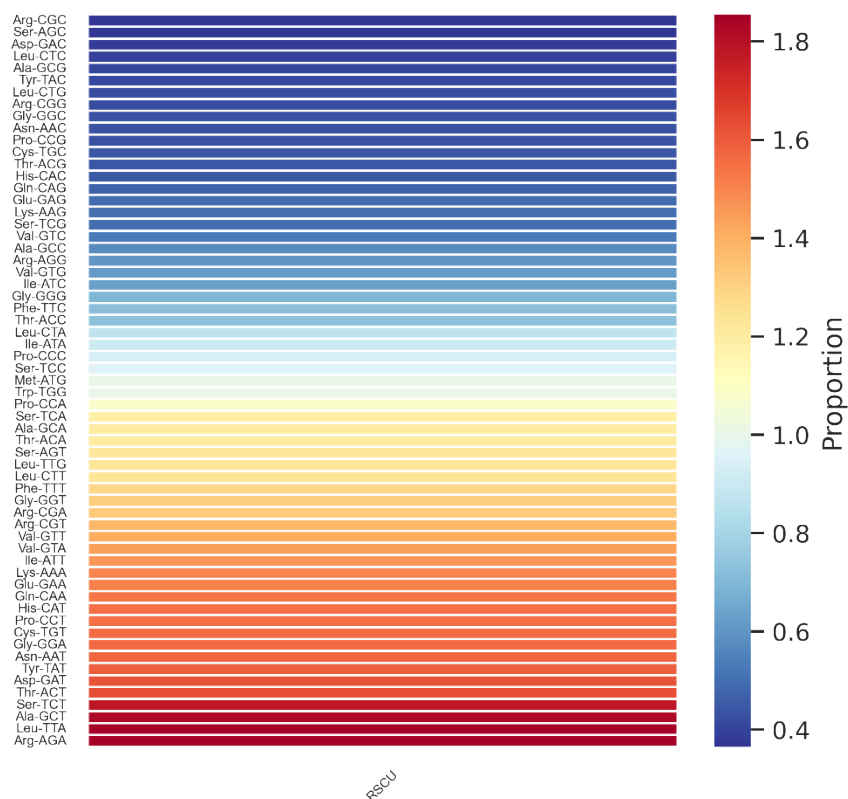


Fig. 3. The figure depicts visualized relative synonymous codon usage in *S. aurea* chloroplast genome as heatmap.

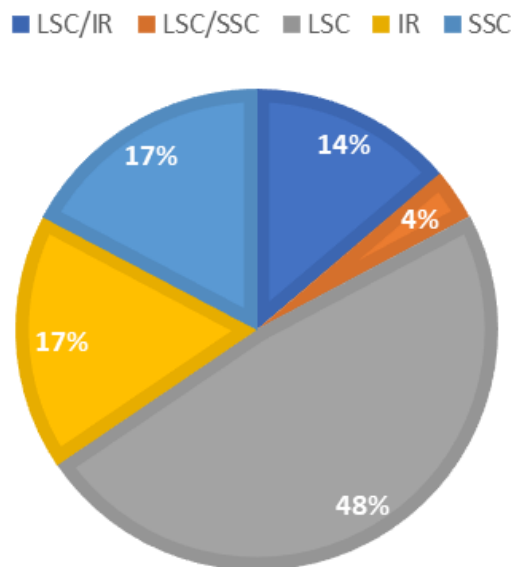


Fig. 4. Percentage location of long repeats in the *S. aurea* cp genomes.

Comparison of cp Genome Structures within Collabieae species

The lengths and gene contents for *S. aurea* and eight other Collabieae cp genomes are available in Table 2. Except for *Cephalantheropsis obcordata* and *Hancockia uniflora*, all other Collabieae species are widely distributed across Peninsular Malaysia. In general, all cp genomes show minimal variation in total length, ranging from 156,036 -158,828 bp. There is also minimal variation in length between LSC, SSC and IR. The total gene contents are conserved for all species.

Table 2. This table shows genomic features of 9 species in Collabieae tribe.

Taxon	Genbank Accession Number	Genome Size (bp)	GC Contents (%)	Gene (Unique)			Region Length (bp)		
				PCG	tRNA	rRNA	LSC	SSC	IR
<i>Spathoglottis aurea</i>	OQ411079	157,957	37.3	78	30	4	86,888	18,125	26,472
<i>Calanthe sylvatica</i>	MK736029	158,828	36.7	78	30	4	87,277	18,535	26,500
<i>Calanthe triplicata</i>	KF753635	158,759	36.7	78	30	4	87,305	18,460	26,497
<i>Calanthe lyroglossa</i>	MN708350	156,036	36.9	78	30	4	85,526	18,254	26,128
<i>Cephalantheropsis obcordata</i>	MN708351	157,918	36.8	78	30	4	86,650	18,420	26,424
<i>Collabium formosanum</i>	MZ618356	157,687	37.2	78	30	4	86,316	18,301	26,535
<i>Hancockia uniflora</i>	OK012601	157,833	37.2	78	30	4	86,345	18,428	26,530
<i>Phaius tankervilleae</i>	MN708349	158,229	37	78	30	4	86,638	18,357	26,617
<i>Tainia acuminata</i>	OL489753	157,603	37.3	78	30	4	86,336	18,129	26,567

* *S. aurea* (OQ411079) data was generated from this study, while other data were downloaded from Genbank.

Gene rearrangements among Collabieae cp genomes are shown in Figure 5. The collinear block lines connecting all the cp genomes demonstrate similar gene arrangement, except for *Collabium formosanum*, in which the genes within the SSC region appear to be inverted.

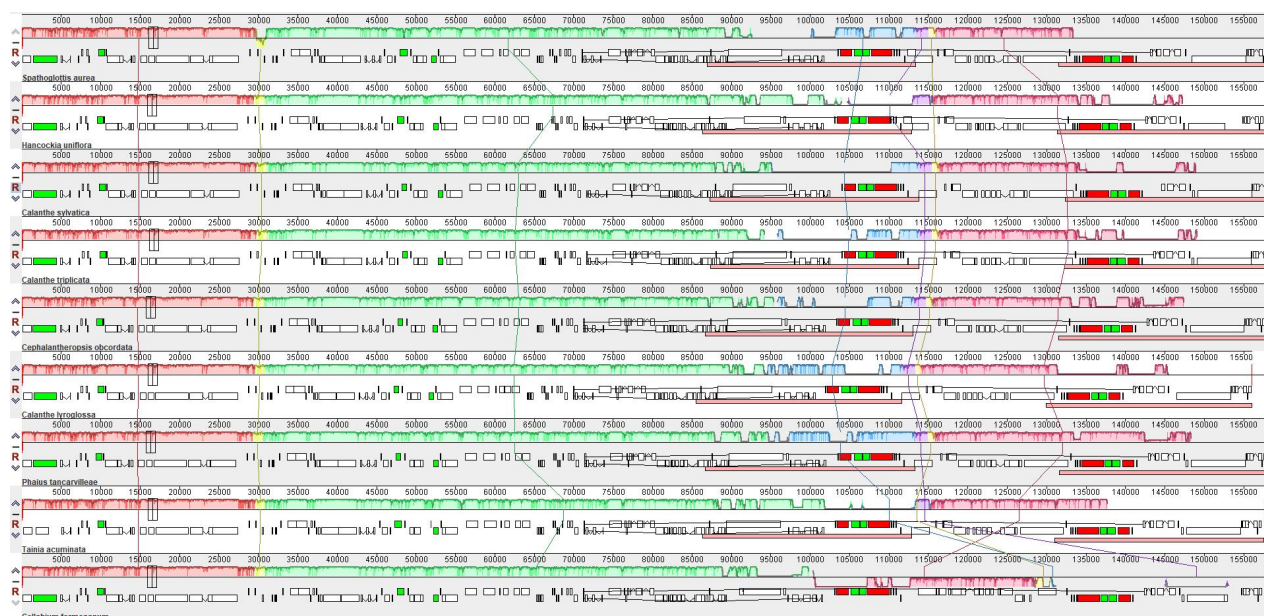


Fig. 5. This figure was generated by progressiveMauve alignment which visualized any genes rearrangements among Collabieae tribe.

Within the eight Collabieae species, there are several variations at the junctions of LSC/IR/SSC. In *S. aurea* and *Calanthe lyroglossa*, the *rpl22* gene is entirely located in the LSC region, at distances of 35 bp and 9 bp, respectively, from the JLB boundary. Conversely, for other species, *rpl22* is located within the JLB boundary, with various lengths (Figure 6). The entire *ycf1* gene is located within the JSB boundary for all species except for *Collabium formosanum*, which is located in the SSC region without overlapping any boundaries. Annotation from *S. aurea* and *Hancockia uniflora* resulted in truncated *ycf1* to be observed in the IRa region near the JSA boundary. The *ycf1* in these two species is part of a duplicated gene in which the complete gene copy was found at the JSB boundary. Likewise, the entire *ndhF* gene for *Collabium formosanum* is located within the SSC region, similar to that in *Hancockia uniflora*. For other species, *ndhF* is slightly expanded from the SSC to the IRa region, thus overlapping within the JSA boundary. For *rps19*, which is in both IR regions, the nucleotide distance from JLA boundaries varies from 59 – 463 bp.

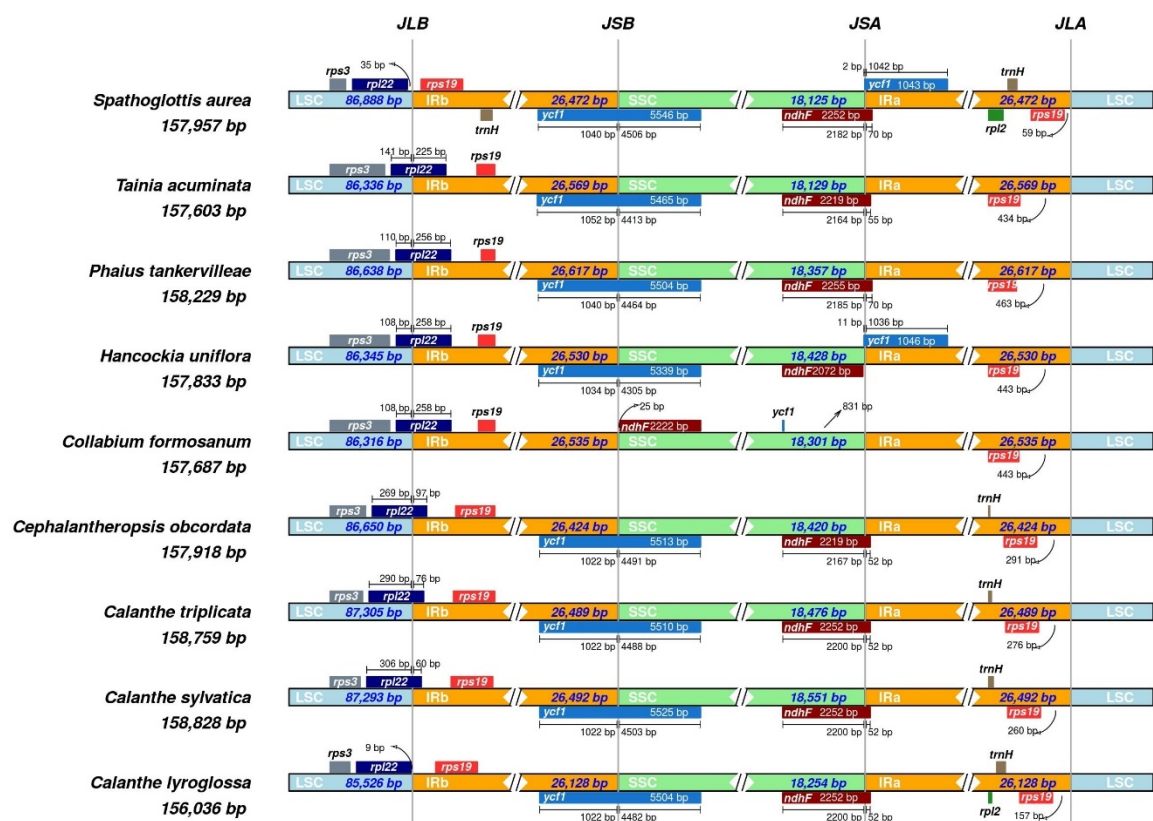


Fig. 6. Comparison of LSC, SSC and IR junctions among 9 Collabieae species. Genes that are presented on top of the track are transcribed anti-clockwise while genes that are presented below of track are transcribed clockwise. The junctions are labelled as JLB (LSC/IRb), JSB (IRb/SSC), JSA (SSC/IRa) and JLA (IRa/LSC).

For nucleotide diversity (π) analysis, *Collabium formosanum* was excluded from the data matrix because of its different SSC orientation. The threshold value for the data matrix is 0.04612; any regions with values higher than this are considered to have high nucleotide diversity. Seven regions demonstrated high nucleotide diversity, including *trnK*-UUU, *rps16* - *trnQ*-UUG, *trnC*-GCA - *trnY*-GUA, *ndhC* - *trnV*-UAC, *atpB* - *rbcL* and *psbB*-*psbT* (Figure 8). Such regions only occur in LSC and not in IR and SSC regions.

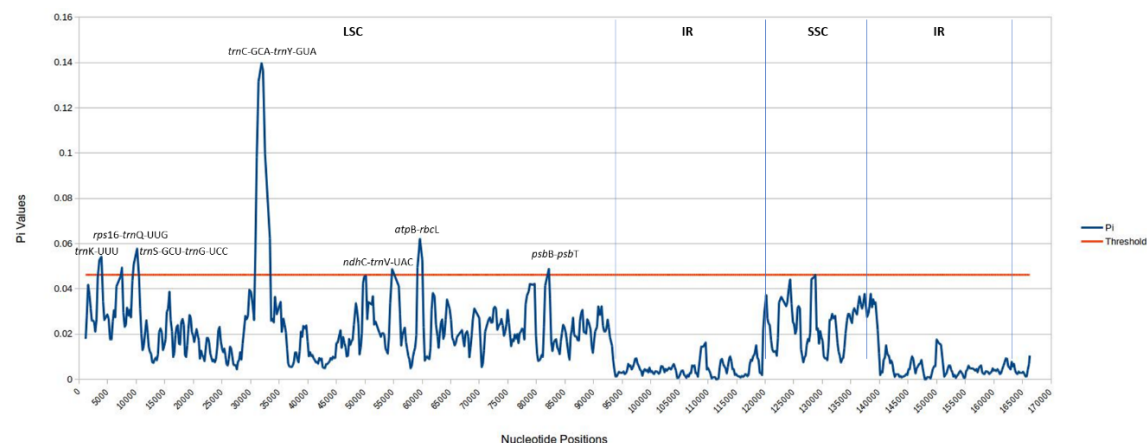


Fig. 7. Nucleotide diversity (Pi) value of *S. aurea* chloroplast genome compared with several Collabieae species.

Phylogenetic Analyses of cp within Collabieae

Further investigation of the phylogenetic position of *S. aurea* among Collabieae species was calculated based on the maximum likelihood (ML) algorithm (Figure 8). The outgroup for the tree is *Dendrobium sinense*, a species from the Malaxideae tribe. Clade A shows *S. aurea* [BP(ML) = 100%] bifurcate farther from other species. Clade B consists of *Collabium formosanum* with *Hancockia uniflora* as a sister clade [BP(ML) = 100%], which appeared to have a closer relationship to species in Clade C [BP(ML) = 100%] than other species involved in this study. All species of *Tainia* genus are clustered in Clade C. Finally, Clade D consists of the paraphyletic *Calanthe* genus along with two closely related genera, *Cephalantheropsis* and *Phaius*.

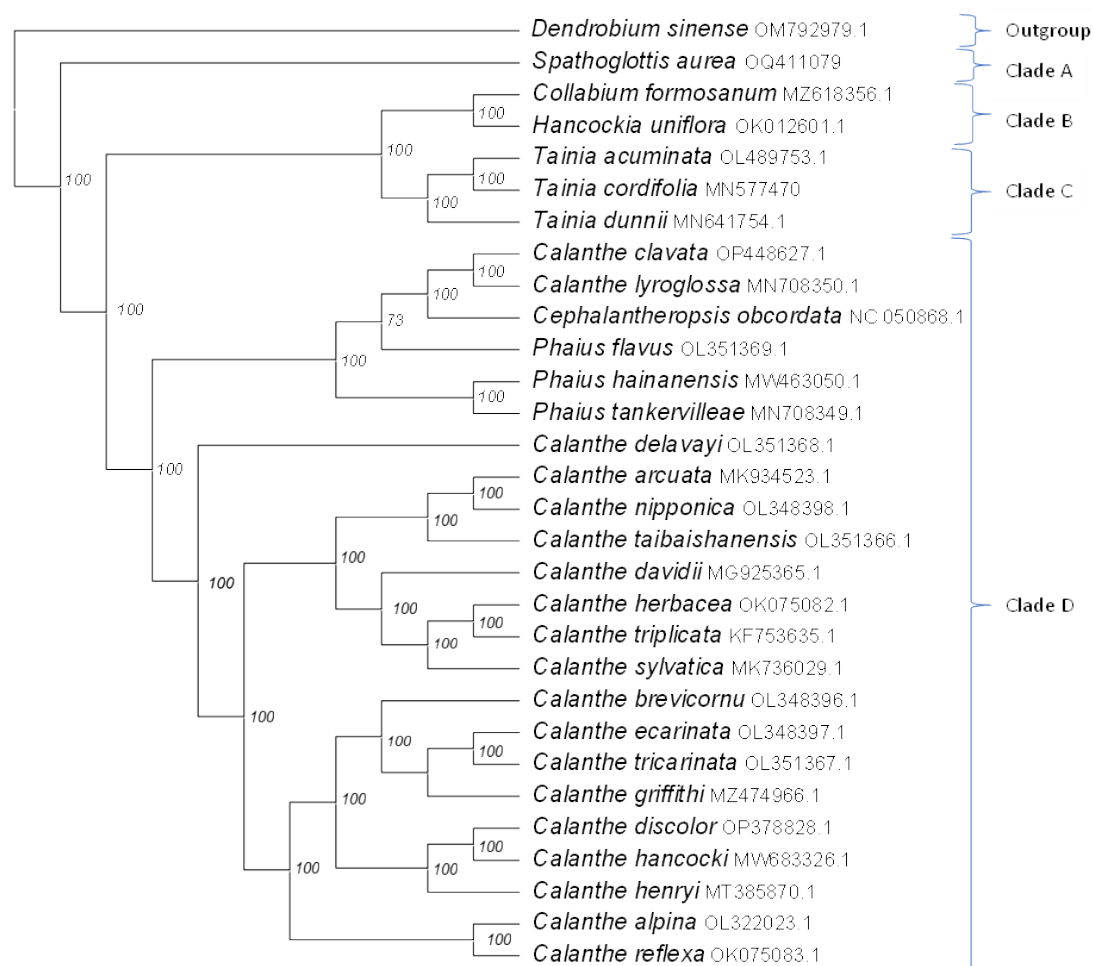


Fig. 8. Phylogenetic tree was built by maximum likelihood (ML) algorithm using 71 PCGs shared by 29 Collabieae species and species outgroup (*Dendrobium sinense*).

Discussion

This study provides the first complete chloroplast genome for the *Spathoglottis* genus. The data will lead to the availability of omics repositories that can support further systematics, evolutionary, population and conservation research (15). Like most other angiosperm chloroplast, such as *Arabidopsis thaliana*, *S. aurea* possessed both cis- and trans-splicing genes. Cis-splicing genes are also referred to as normal gene splicing, whereby the joined exons originated from the same transcript. In contrast, trans-splicing genes require the joining of exons from different gene locations prior to mRNA processing (37). Other than the 13 PCGs in Figure 2, there are also six tRNA (*trnG*-UCC, *trnL*-UAA, *trnV*-UAC, *trnI*-GAU, *trnA*-UGC, *trnV*-GAC) genes and an rRNA (*rrn23*) gene that are cis-splicing genes in the *S. aurea* cp genome. *rps12* is the most reported trans-splicing gene in the chloroplast genome and can exist in several configurations depending on the species (25). The configurations are based on the gene locations in the cp genome.

The development of microsatellites or SSR markers are very useful for populations and genetic relationship studies. Although SSR markers developed from the nuclear region give comprehensive information about both maternal and paternal parents, their high rate of sequence evolution makes them unsuitable for phylogenetic analysis of different taxa (38). SSR markers developed from the chloroplast region could be complementary here, because of this region's highly conserved nature compared to that of the nuclear region (39).

Compared to other Collabieae species, the *S. aurea* cp genome shows a highly conserved structure without any gene loss. The quadripartite structure and gene compositions are consistent with previously reported Collabieae species (9,10,40). In Figure 5, gene inversion was observed in the *Collabium formosanum* cp genome because of its different orientation submitted in the GenBank (41). It is the nature of cp genomes to harbour two different structure orientations based on the SSC region. Thus, there is no gene rearrangement between Collabieae species in this study. Similarly, in Figure 6, the gene *ndhF* seems to be positioned differently from other species because of a different SSC orientation. Nevertheless, in *Collabium formosanum* there is no *ycf1* gene available in the IR region because it is fully located inside the SSC region.

Nucleotide diversity (Pi) analysis of *S. aurea* revealed several mutational hotspots compared to seven other Collabieae species. Species from earlier analysis were used, except for *Collabium formosanum*. Different orientations of the SSC regions might lead to false interpretations. In Figure 7, seven regions are shown, which are mostly intergenic regions except for the *trnC*-GCA - *trnY*-GUA region which also includes *psbM*, *petN* and *trnD*-GUC as highly variable regions. These regions are useful for the development of molecular markers to resolve several taxonomic debates reported by previous studies (1,42,43).

The phylogenetic tree in Figure 8 revealed similar results to the molecular study by Xiang (44), which use four chloroplast genes (*matK*, *psaB*, *rbcL* and *trnH-psbA*) to elucidate the relationship among Collabieae species. In the present study, *S. aurea* shows a further relationship with other species (Clade A); in particular, *Cephalantheropsis* and *Phaius* also reside in the same clade as *Calanthe*. Despite several differences in morphological characteristics, a number of molecular studies (9,44) including the present study have consistently shown that these two genera are positioned within the *Calanthe* clade. For the *Phaius* species, the morphological characteristic differences are minimal compared to those of the *Calanthe* species. However, species under the *Cephalantheropsis* genus showed significant differences in morphological characteristics compared to *Calanthe* and *Phaius*. Chase proposed to merge these two genera (*Cephalantheropsis* and *Phaius*) under *Calanthe* (16,45), but there is ongoing debate about this merger and it has not been universally accepted.

Conclusion

This study characterized the complete chloroplast genome of *S. aurea*; one of the commonly known species used in landscaping and to develop artificial hybrids. The results provide the first genomic resource of the *Spathoglottis* genus. Data such as SSR locations will be useful for future assessment of *S. aurea* populations and other *Spathoglottis* species. The evolutionary relationship between *S. aurea* and other Collabieae species based on shared PCGs was also revealed to be consistent with previous studies, including the highly debated genera *Phaius* and *Cephalantheropsis* in Orchidaceae.

Supporting information

S1 Fig. Coverage graph. Coverage depth of nucleotide position of *S. aurea* chloroplast genome

S1 Table. Codon table. Codon usage table *S. aurea* chloroplast genome

S2 Table. MISA file. Original MISA file for SSR detection in *S. aurea* chloroplast genome

S3 Table. SSR table. Details on SSR location in the *S. aurea* chloroplast genome

Acknowledgement

The study was funded by the USM Short Term Grant of assignment No. 304/PBI- OLOGI/6315549 awarded to the third author and the corresponding author. All funders provided financial supports for the present work but did not contribute any additional role in in the research design, data collections and analysis, and preparation of the manuscript.

Author Contributions

Conceptualization: Ahmad Sofiman Othman, Nurul Shakina Mohd Talkah

Data curation: Nurul Shakina Mohd Talkah

Formal analysis: Nurul Shakina Mohd Talkah

Funding acquisition: Farah Alia Nordin, Ahmad Sofiman Othman

Project administration: Ahmad Sofiman Othman

Resources: Farah Alia Nordin, Jasim Haider Mahmod

Software: Nurul Shakina Mohd Talkah

Writing - original draft: Nurul Shakina Mohd Talkah, Jasim Haider Mahmod

Writing - review & editing: Ahmad Sofiman Othman, Farah Alia Nordin

References

1. Wei N, Pérez-Escobar OA, Musili PM, Huang W-C, Yang J-B, Hu A-Q, et al. Plastome Evolution in the Hyperdiverse Genus *Euphorbia* (Euphorbiaceae) Using Phylogenomic and Comparative Analyses: Large-Scale Expansion and Contraction of the Inverted Repeat Region [Internet]. Vol. 12, *Frontiers in Plant Science* . 2021. Available from: <https://www.frontiersin.org/articles/10.3389/fpls.2021.712064>
2. Peredo EL, King UM, Les DH. The plastid genome of *Najas flexilis*: adaptation to submersed environments is accompanied by the complete loss of the NDH complex in an aquatic angiosperm. *PLoS One*. 2013;8(7):e68591.
3. Olmstead RG, Palmer JD. Chloroplast DNA systematics: a review of methods and data analysis. *Am J Bot*. 1994;81(9):1205–24.
4. Clegg MT, Zurawski G. Chloroplast DNA and the study of plant phylogeny: present

- status and future prospects. *Mol Syst plants*. 1992;1–13.
5. Hollingsworth PM, Graham SW, Little DP. Choosing and using a plant DNA barcode. *PLoS One*. 2011;6(5):e19254.
6. Butzin F. Die Namen der supragenerischen Einheiten der Orchidaceae (The Names of the Suprageneric Units of Orchidaceae). *Willdenowia*. 1971;301–40.
7. Chase MW, Cameron KM, Freudenstein J V, Pridgeon AM, Salazar G, Van den Berg C, et al. An updated classification of Orchidaceae. *Bot J Linn Soc*. 2015;177(2):151–74.
8. Zhong H, Shen L-M, Liu H-P, Liu Z-J, Wu S-S, Zhai J-W. The complete chloroplast genome of *Calanthe arcuata*, an endemic terrestrial orchid in China. *Mitochondrial DNA Part B*. 2019;4(2):2629–30.
9. Nanjala C, Wanga VO, Odago W, Mutinda ES, Waswa EN, Oulo MA, et al. Plastome structure of 8 *Calanthe* species (Orchidaceae): comparative genomics, phylogenetic analysis. *BMC Plant Biol*. 2022;22(1):1–22.
10. Miao L-Y, Hu C, Huang W-C, Jiang K. Chloroplast genome structure and phylogenetic position of *Calanthe sylvatica* (Thou.) Lindl.(Orchidaceae). *Mitochondrial DNA Part B*. 2019;4(2):2625–6.
11. Zhang S-D, Wang Q, Du M-M, Ling L-Z. Characterization of the chloroplast genome of *Calanthe henryi* (Epidendroideae; Orchidaceae). *Mitochondrial DNA Part B*. 2020;5(3):2273–5.
12. Zheng Q-D, Zhou J, Ma S-H, Chen M-K, Xie T-X, Chen J, et al. The complete chloroplast genome sequence of *Tainia cordifolia* (Orchidaceae). *Mitochondrial DNA Part B*. 2020;5(1):1–2.
13. Xie T-X, Yu X, Zheng Q-D, Ma S-H, Liu Z-J, Ai Y. The complete chloroplast genome of *Tainia dunnii* (Orchidaceae): genome structure and evolution. *Mitochondrial DNA Part B*. 2020;5(1):3–4.
14. Ou X, Qin Y, Zhang L, Sun T, Qin X. The complete chloroplast genome of the newly recorded species *Tainia acuminata* Averyanov (Orchidaceae) from China. *Mitochondrial DNA Part B*. 2022;7(5):884–5.
15. Nordin FA, Saibeh K, Go R, Mangsor KNA, Othman AS. Molecular Phylogenetics of the Orchid Genus *Spathoglottis* (Orchidaceae: Collabieae) in Peninsular Malaysia and Borneo. *Forests*. 2022;13(12):2079.
16. Chase MW, Christenhusz MJM, Schuiteman A. Expansion of *Calanthe* to include the species of *Cephalantheropsis*, *Gastrorchis* and *Phaius* (Collabieae; Orchidaceae). *Phytotaxa*. 2020;472(2):159–68.
17. Pridgeon AM, Cribb PJ, Chase MW, Rasmussen FN. *Genera Orchidacearum*. Epidendroideae Vol. 4 (part 1). Oxford University Press. New York, USA; 2005.
18. Nordin FA, Raffi A, Go R, Seok Yien Yong C, Saibeh K, Othman AS. Morphological Systematics of *Spathoglottis* Blume (Orchidaceae: Collabieae) in Peninsular Malaysia and Borneo. Vol. 14, *Forests*. 2023.
19. Doyle JJ, Doyle JL. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. 1987.
20. Jin J-J, Yu W-B, Yang J-B, Song Y, dePamphilis CW, Yi T-S, et al. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol* [Internet]. 2020;21(1):241. Available from: <https://doi.org/10.1186/s13059-020-02154-5>
21. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics*. 2009;25(14):1754–60.
22. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–9.

23. Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, et al. GeSeq—versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* 2017;45(W1):W6–11.
24. Boratyn GM, Camacho C, Cooper PS, Coulouris G, Fong A, Ma N, et al. BLAST: a more efficient report with usability improvements. *Nucleic Acids Res [Internet]*. 2013 Jul 1;41(W1):W29–33. Available from: <https://doi.org/10.1093/nar/gkt282>
25. Liu S, Ni Y, Li J, Zhang X, Yang H, Chen H, et al. CPGView: A package for visualizing detailed chloroplast genome structures. *Mol Ecol Resour.* 2023;
26. Beier S, Thiel T, Münch T, Scholz U, Mascher M. MISA-web: a web server for microsatellite prediction. *Bioinformatics.* 2017;33(16):2583–5.
27. Kurtz S, Choudhuri J V, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* 2001;29(22):4633–42.
28. Stothard P. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *Biotechniques.* 2000;28(6):1102–4.
29. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol.* 2018;35(6):1547.
30. Darling ACE, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* 2004;14(7):1394–403.
31. Amiryousefi A, Hyvönen J, Poczai P. IRscope: an online program to visualize the junction sites of chloroplast genomes. *Bioinformatics [Internet]*. 2018 Sep 1;34(17):3030–1. Available from: <https://doi.org/10.1093/bioinformatics/bty220>
32. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 2013;30(4):772–80.
33. Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, et al. DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol Biol Evol.* 2017;34(12):3299–302.
34. Celiński K, Kijak H, Wiland-Szymańska J. Complete chloroplast genome sequence and phylogenetic inference of the Canary Islands Dragon Tree (*Dracaena draco* L.). *Forests.* 2020;11(3):309.
35. Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermini LS. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods.* 2017;14(6):587–9.
36. Trifinopoulos J, Nguyen L-T, von Haeseler A, Minh BQ. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res.* 2016;44(W1):W232–5.
37. Lasda EL, Blumenthal T. Trans-splicing. *Wiley Interdiscip Rev RNA.* 2011;2(3):417–34.
38. Powell W, Morgante M, Andre C, Hanafey M, Vogel J, Tingey S, et al. The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. *Mol Breed.* 1996;2:225–38.
39. Huang J, Yang X, Zhang C, Yin X, Liu S, Li X. Development of chloroplast microsatellite markers and analysis of chloroplast diversity in Chinese jujube (*Ziziphus jujuba* Mill.) and wild jujube (*Ziziphus acidojujuba* Mill.). *PLoS One.* 2015;10(9):e0134519.
40. Yang J, Li D, Li H. Highly effective sequencing whole chloroplast genomes of angiosperms by nine novel universal primer pairs. *Mol Ecol Resour.* 2014;14(5):1024–31.
41. Turudić A, Liber Z, Grdiša M, Jakše J, Varga F, Šatović Z. Towards the well-tempered chloroplast DNA sequences. *Plants.* 2021;10(7):1360.

42. Zhou T, Wang J, Jia Y, Li W, Xu F, Wang X. Comparative Chloroplast Genome Analyses of Species in *Gentiana* section *Cruciata* (Gentianaceae) and the Development of Authentication Markers. *Int J Mol Sci* [Internet]. 2018 Jul 5;19(7):1962. Available from: <https://pubmed.ncbi.nlm.nih.gov/29976857>
43. Yang J, Vázquez L, Chen X, Li H, Zhang H, Liu Z, et al. Development of chloroplast and nuclear DNA markers for Chinese oaks (*Quercus* subgenus *Quercus*) and assessment of their utility as DNA barcodes. *Front Plant Sci*. 2017;8:816.
44. Xiang X-G, Jin W-T, Li D-Z, Schuiteman A, Huang W-C, Li J-W, et al. Phylogenetics of tribe Collabieae (Orchidaceae, Epidendroideae) based on four chloroplast genes with morphological appraisal. *PLoS One*. 2014;9(1):e87625.
45. Chase MW, Christenhusz MJM, Schuiteman A. (2782) Proposal to conserve *Calanthe*, nom. cons., against the additional names *Phaius*, *Cyanorkis*, and *Gastorkis* (Orchidaceae, Collabieae). *Taxon*. 2020;69(6):1364–5.