Revista de Biología Tropical, ISSN: 2215-2075, Vol. 72: e56423, enero-diciembre 2024 (Publicado May. 21, 2024)

BIOTECHNOLOGY

REVISTA DE
Biología Tropical

https://doi.org/10.15517/rev.biol.trop..v72i1.56423

Complete chloroplast genome of the Jewel Orchid, *Anoectochilus formosanus* (Orchidaceae) and its relatives

Minh Phuong Nguyen[1]; https://orcid.org/0000-0003-1768-5555
Thi Huong Trinh[1]; https://orcid.org/0000-0002-5940-6588
Khuong Duy Dao[2, 3]; https://orcid.org/0009-0005-1607-2871
Phuc Loi Luu[2, 4]; https://orcid.org/0000-0001-8045-718X
Viet The Ho[1]*; https://orcid.org/0000-0003-4863-0530

1 Faculty of Biology and Environment, Ho Chi Minh City University of Industry and Trade, 140 Le Trong Tan, Tan Phu district, Ho Chi Minh City 700000, Vietnam; phuongnguyen@huit.edu.vn, trinhth@huit.edu.vn, thehv@huit.edu.vn (*Correspondence)
2 Tam Anh Research Institute (TamRI), 2B Pho Quang Street, Ward 2, Tan Binh District, Ho Chi Minh City 700000, Vietnam; loilp@tamri.vn, duydk@tamri.vn
3 Faculty of Biology & Biotechnology, The University of Science, Viet Nam National University Ho Chi Minh City, 227 Nguyen Van Cu, Ward 4, District 5, Ho Chi Minh City 700000, Vietnam; duydk@tamri.vn
4 Mathematics Department, Faculty of Fundamental Sciences, University of Medicine and Pharmacy at Ho Chi Minh City (UMP), 217 Hong Bang Street, Ward 11, District 5, Ho Chi Minh City 700000, Vietnam; luuphucloi@ump.edu.vn

Received 31-VIII-2023. Corrected 06-II-2024. Accepted 16-V-2024.

ABSTRACT

Introduction: *Anoectochilus formosanus* is a highly valuable herb known for its efficacy in treating a wide range of diseases. However, the current methods used to differentiate this species from others within the same genus are not effective due to the high similarity in morphological characteristics and DNA barcode sequences among these species.

Objective: To characterize the chloroplast (cp) genome to distinguish *A. formosanus* at species or isolation levels.

Methods: The complete cp genome was sequenced using next-generation sequencing technology, annotated, and compared with published cp genomes of various species within the *Anoectochilus* genus.

Results: The complete cp genome of *A. formosanus* is 152 658 bp in size, consisting of a large and small copy of 82 692 bp and 17 346 bp, respectively, separated by inverted repeats of 26 310 bp. Within the cp genome, there are a total of 141 genes, including 92 protein-coding genes, 10 rRNA genes, and 39 tRNA genes. This genome contains a total of 80 simple sequence repeats, with 50 long repeats. Through phylogenetic analysis, a close relationship was observed between *A. formosanus* in Vietnam and *A. formosanus* samples from China (NC_061756.1). However, genomic comparisons highlighted differences between the two cp genomes, specifically in their reverse repeat sequences.

Conclusion: These findings reveal distinct variations in the cp genome of *A. formosanus* in Vietnam, offering valuable insights into the taxonomy, plant identification, breeding, and conservation programs related to this herb in Vietnam.

Key words: *Anoectochilus formosanus*; chloroplast genome; medicinal plant; next generation sequencing; simple sequence repeat.

RESUMEN

Genoma completo del cloroplasto de la orquídea joya, *Anoectochilus formosanus* (Orchidaceae) y sus afines

Introducción: *Anoectochilus formosanus* es una orquídea herbácea muy valiosa, conocida por su eficacia en el tratamiento de una amplia gama de enfermedades. Sin embargo, los métodos actuales utilizados para diferenciar esta especie de otras dentro del mismo género no son efectivos debido a la gran similitud en las características morfológicas y las secuencias de códigos de barras de ADN entre estas especies.
Objetivo: Caracterizar el genoma del cloroplasto (cp) para distinguir *A. formosanus* a nivel de especie o de aislamiento.
Métodos: El genoma completo del cp se secuenció utilizando tecnología de secuenciación de nueva generación, se anotó y se comparó con los genomas del cp publicados de varias especies del género *Anoectochilus*.
Resultados: El genoma del cp completo de *A. formosanus* tiene un tamaño de 152 658 pb y consta de una copia grande y pequeña de 82 692 pb y 17 346 pb, respectivamente, separadas por repeticiones invertidas de 26 310 pb. Dentro del genoma del cp, hay un total de 141 genes, incluidos 92 genes codificadores de proteínas, 10 genes de ARNr y 39 genes de ARNt. Este genoma contiene un total de 80 repeticiones de secuencia simple, con 50 repeticiones largas. Mediante análisis filogenético, se observó una estrecha relación entre *A. formosanus* de Vietnam y muestras de *A. formosanus* de China (NC_061756.1). Sin embargo, las comparaciones genómicas resaltaron diferencias entre los dos genomas del cp, específicamente en sus secuencias de repetición invertida.
Conclusión: Estos hallazgos revelan distintas variaciones en el genoma del cp de *A. formosanus* en Vietnam, lo que ofrece información valiosa sobre la taxonomía, la identificación de plantas, la reproducción y los programas de conservación relacionados con esta hierba en Vietnam.

Palabras clave: *Anoectochilus formosanus*; genoma del cloroplasto; planta medicinal; secuenciación de próxima generación; repetición de secuencia simple.

INTRODUCTION

Anoectochilus formosanus is a medicinal plant that belongs to the Orchidaceae family (Shiau et al., 2002). It has been utilized for centuries as a traditional medicine due to its powerful healing properties. In recent times, there has been a growing interest in scientific research to explore the pharmacological potential of this plant. Several studies have demonstrated that extracts from *A. formosanus* exhibit substantial antioxidant activity and effectively reduce the production of inflammatory (Ho et al., 2018; Lin et al., 1993; Shiau et al., 2002). Moreover, the extract has shown promising anti-hyperglycemic effects in diabetic rats by significantly reducing blood sugar levels and improving insulin resistance, indicating its potential therapeutic application in diabetes treatment in mice (Shih et al., 2002; Tang et al., 2018). Additionally, it possesses various pharmacological activities, such as antioxidant, anti-inflammatory, anticancer, and neuroprotective effects (Nguyen et al., 2023b; Wang et al., 2002). The exceptional medicinal value of this herb can be attributed to its chemical composition, which includes phenolic acids, flavonoids, diarylpentanoid, kinsenone, and polysaccharides (Chiang & Lin, 2017; Wang et al., 2002; Xu et al., 2022).

A. formosanus is highly sought after but faces limited supply, raising concerns among conservationists and researchers regarding overharvesting. The species is on the brink of extinction due to multiple factors, including habitat destruction, overexploitation, illegal trade, and climate change (Jiang et al., 2015, Kumar & Gale, 2020; Ma et al., 2010; Zhang et al., 2013). A previous study using ISSR and AFLP markers revealed that *A. formosanus* populations exhibit relatively low genetic diversity, indicating their susceptibility to environmental changes and overexploitation (Lin et al., 2007). These findings underscore the urgent need for conservation efforts and the adoption of sustainable practices to safeguard *A. formosanus* populations from further depletion and the risk of extinction.

Traditionally, morphological identification has been utilized for the conservation-oriented identification of jewel orchids. However, this approach has limitations, particularly when closely related species share similar morphological characteristics (Suetsugu et al., 2022; Tran et al., 2022). Certain orchid species exhibit comparable leaf colors and patterns, creating challenges for differentiation. Furthermore, some orchid species display considerable variation in morphological features, including leaf shape and size, which further complicates identification (Huynh et al. 2019; Nguyen et al., 2023a). Due to the significant overlap in their morphological traits, the reliability of relying solely on morphological identification for jewel orchids, especially for closely related species with similar morphological characteristics, is questionable (Bhattacharjee & Chowdhery, 2013; Hu et al., 2016; Ong & Lee, 2019; Suetsugu et al., 2022).

In recent years, DNA barcodes have gained significant popularity for the identification and classification of various plant species, including orchids. However, this method has certain limitations, especially when dealing with closely related species that share similar DNA sequences. Ho et al. (2021) discovered that commonly used DNA barcoding markers, such as *rbc*L and *mat*K, have limited and variable discriminatory power in distinguishing closely related jewel orchid species. Furthermore, some orchid species exhibit highly variable DNA sequences, which can further complicate their identification through DNA barcoding (Huynh et al., 2019). Additionally, there is considerable overlap in the DNA sequences of certain jewel orchid species, rendering DNA barcoding an unreliable method (Chen & Shiau, 2015; Ho et al., 2021; Zhang et al., 2019).

The accessibility and affordability of next-generation sequencing (NGS) technology have revolutionized the characterization of complete chloroplast (cp) genomes, proving to be an effective method for identifying and classifying orchids (Konhar et al., 2019; Tang et al., 2021; Yang et at., 2013). The cp genome, inherited maternally from the parent, evolves slowly, making it a valuable tool for studying molecular evolution, population genetics, phylogenetics, botany, and genomic evolution (Yang et at., 2013). Ho and colleagues identified the differences in cp genomes of *Ludisia discolor* accession collected in Vietnam with three assessions from China (Ho et al., 2023). These findings emphasize the potential of NGS technology in identifying and classifying jade orchids based on their cp genome. To our best knowledge, only one cp genome of *A. formosanus* with China origin has been published with NC_061756.1 accession number in the National Center for Biotechnology Information (NCBI) genbank. Although this medicinal plant is considered as endemic of Vietnam but not such information available. In our specific study, we utilized NGS to sequence the cp genome of *A. formosanus* samples collected in Vietnam, comparing the results with published sequences to identify distinctive genomic features. This information holds valuable implications for the taxonomy, botanical identification, breeding, and conservation programs of *A. formosanus* in Vietnam.

MATERIALS AND METHODS

Sample collection, DNA extraction, library construction, and sequencing: Samples of *A. formosanus* were provided by the Biotechnology Center of Ho Chi Minh City (HCMBIOTECH), the voucher sample is kept at this center for conservation. Total DNA was extracted from fresh leaves using the Isolate II Plant DNA Kit (Bioline, UK). DNA quality and quantity were determined by 1 % gel electrophoresis and Nanodrop, respectively (Thermo-Scientific, Delaware, USA).

100-1 000 ng DNA that had undergone quality control, was fragmented using acoustic disruption with Covaris S220, followed by final repair, dA tailing, adapter ligation, and purification. The purified DNA was then selected for the appropriate size before being PCR amplified for library construction. Preliminary quantification and library dilution were carried out using Qubit3.0, followed by the use of Agilent 2 100 to determine the insert size and nucleic acid concentration of the resulting library sample. The effective concentration of each sample library in the mixture was determined by qPCR (Realtime PCR 7 500-Applied Biosystem, USA) before sequencing to ensure the accuracy of the sample concentration and the reliability of the sequencing data.

Base calling was accomplished using the RTA software integrated with the Illumina Novaseq 6 000 sequencer, which converted the four fluorescence signals obtained from the Charge-Coupled Device (CCD) to binary bcl data in real-time. The bcl data were then transformed to a fastq file using bcl2fastq (v.2.17), which is part of the software package provided by Illumina. Concurrent demultiplexing of the data was carried out based on the index information. Primary analysis was conducted using the built-in High-Content Screening (HCS) sequencer software (Ogier & Dorval, 2012) to determine whether the read data passed the quality filter, based on the signal quality of the first 25 cycles. The quality of the raw reads was initially assessed using FastQC in the Galaxy portal (The Galaxy Community, 2022) and then submitted to the Sequence Read Archive (SRA) database in NCBI (2009) under the PRJNA982609 project.

Chloroplast genome assembly and annotation: To eliminate adapter content, the sequence data were processed using Trim Galore (v.0.6.7) on Galaxy (Li et al., 2023), which is a tool built into The Galaxy Server. Reference-based assembly was performed using the HISAT2 (v.2.2.1) tool, aligning the data against the *A. formosanus* reference sequence with NCBI accession number NC_061756 to generate BAM files. The Pilon (v.1.21) tool was then employed with the assembly and mapping information to identify and correct any potential problems with the assembly and to generate a FASTA file for further analysis. To annotate and localize protein-coding genes, rRNA, and tRNA in the cp genome, the Chloroplot program (Zheng et al., 2020) was utilized.

Comparative analysis among species in *Anoectochilus* genus: To align the cp genome sequences, the MAFFT (v.7) program (Katoh & Standley, 2013) was employed using the parameters according to Katoh et al. (2019). The resulting alignment was then used to identify DNA polymorphisms, and nucleotide diversity (Pi) was calculated using DnaSP (v.6.12.03) software (Rozas et al., 2017). The genetic differences between cp genomes were determined using the Kimura two-parameter algorithm of MEGA X software, based on the *P*-value (P distance) calculated from the evolutionary distance between sequences.

The comparison of cp genomes among related species was carried out using the Shuffle-LAGAN mode of the VISTA program (Frazer et al., 2004). The junctions of LSC/IRB/SSC/IRA were visualized using IRscope (Amiryousefi et al., 2018) based on the cp genome annotations of these related species available in Genbank. The MIcroSAttelite (MISA) identification tool was employed to identify SSR motifs, with ten repeat parameters for mononucleotides, six for dinucleotides, five for trinucleotides, four for tetra-nucleotides, and three for penta- and hexa-nucleotides, as described by Beier et al. (2017). The REPuter software was used to determine long repeat regions with repeat size ≥ 30 bp and a minimum of 90 % identity, identifying four types of repeats, namely forward (F), reverse (R), complement (C), and retrograde (P), as reported by Kurtz et al. (2001).

Phylogenetic analysis: The NCBI Genbank (NCBI, 1982) was used to retrieve additional complete cp genome sequences of various species of jewel orchid from the *Anoectochilus* genus. Unverified sequences were excluded from the analysis, and only one sequence was randomly chosen for further analysis in cases where multiple sequences were available for a particular species. The following eight complete cp genome sequences were obtained: NC_03895.1 *A. emeiensis*; NC_066958 *A. burmannicus*; NC_061758 *A. roxburghii*; NC_061756 *A. formosanus*; NC_054353 .1 *A. zhejiangensis*; MW589501.1 *A. hainanensis*; MW589500.1 *A. chapaensis*; and NC_033895.1 *A. emeiensis*.

The MAFFT alignment were utilized to establish the phylogenetic relationships between the cp genomes. A phylogenetic tree of nine cp genomes was constructed using the Neighbour Joining (NJ) method and Maximum Likelihood (ML) methods representing distance and discrete character methods (Kang et al., 2017) with 5 000 bootstrap replicates by using MEGA X software. The *Dendrobium sinense* cp genome (OM792979.1), a common ornamental orchid, was used as an outgroup. The Kimura 2-parameter nucleotide substitution model, which is commonly employed to estimate genetic differences resulting from nucleotide substitutions (Nishimaki & Sato, 2019), was applied to the phylogenetic trees.

RESULTS

Genome, sequence assembly, and features of chloroplast: A total of 6.9 GB of 150 bp paired-end data was generated, resulting in 19 211 194 reads with a Phred scores of 95.26 % of reads are greater than Q20. The GC content of the plastome was approximately 37 %. Upon assembly, the cp genome map exhibited a conserved circular structure with a total length of 152 658 bp. The genome comprised four distinct parts, including a Large Single Copy (LSC) region spanning 82 692 bp, a Small Single Copy (SSC) region spanning 17 346 bp, and two Inverted Repeat (IR) regions spanning 26 310 bp each. These IR regions were separated by the LSC and SSC regions (Fig. 1).

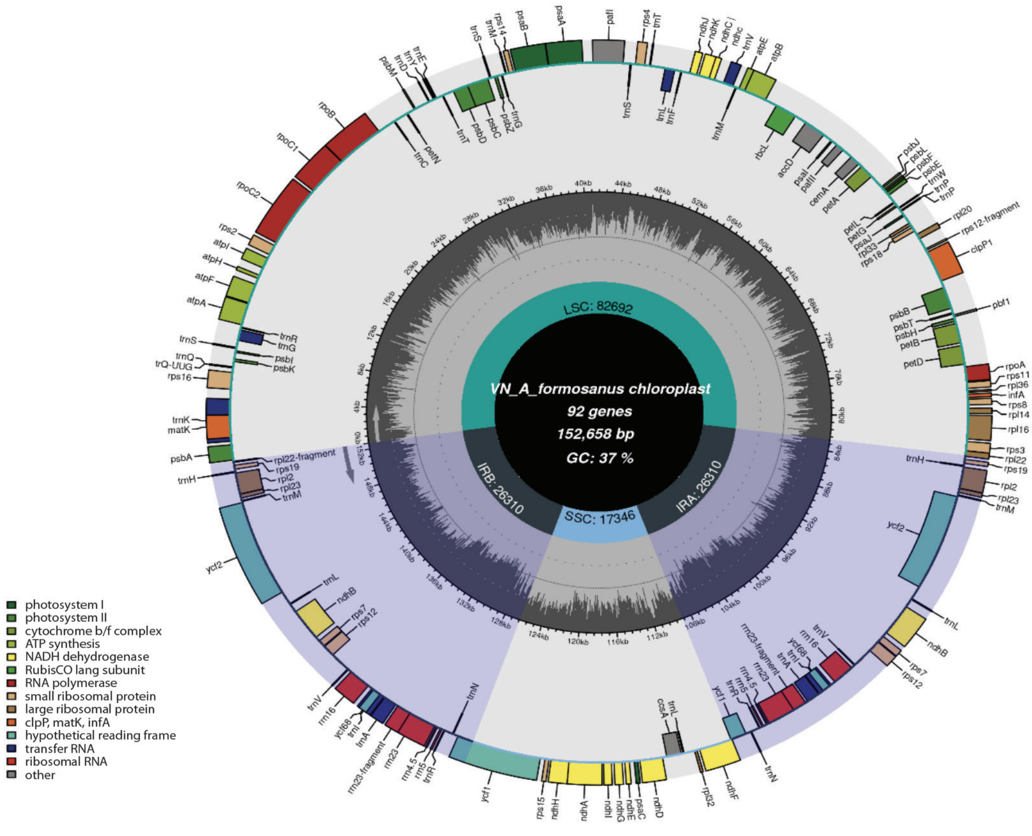

Fig. 1. The cp genome map of Vietnam *A. formosanus,* generated with Chloroplot program, displays the genes transcribed in clockwise and counterclockwise directions, depicted outside and inside of the circle, respectively. The LSC, SSC, IRA, and IRB are labeled as the primary parts of the cp genome. The inner circle's dark and light grey colors represent the GC and AT content, respectively.

Sequence annotation and comparison of cp genomes: A total of 142 genes were annotated in the obtained cp genome, including 92 protein-coding genes, 10 rRNA genes, and 39 tRNA genes. To compare the gene composition of the cp genome, we examined eight other *Anoectochilus* species, whose data had been previously published on NCBI and are presented in Table 1. The cp genome of *A. formosanus* exhibited slight differences compared to the other cp genomes, particularly in the number of coding genes and tRNA genes. Surprisingly, several variations were observed between the *A. formosanus* specimen from Vietnam and the one from China (NC_061765.1).

The divergence among the nine cp genomes ranged from 0.000 to 0.005, as shown in Table 2. The Dnasp program identified a total of 1 464 polymorphic sites from the aligned cp sequences of the *Anoectochilus* genus. The nucleotide diversity value (Pi) was calculated to be 0.0036.

Repeat structure and Simple sequence repeats: Among the nine jewel orchid species, a total of 657 SSRs were detected, ranging from 68 SSRs in *A. calcareus* to 80 SSRs in VN_*A_formosanus*, with an average of approximately 73 SSRs per cp genome (Fig. 2). Six types of SSR motifs were identified: A, T, C, AT, TA, and TTC. The most abundant mononucleotide motifs were T and A, accounting for 64.2 % (422 SSRs) and 28.0 % (184 SSRs), respectively. Interestingly, only one PolyC motif was present, while no polyG motif was detected in the genome. Short polyA and polyT repeats are commonly observed as SSRs in cp genomes, whereas polyG or polyC repeats are rare (Lei et al., 2016). Additionally, dinucleotide motifs (AT and TA) and a trinucleotide motif (TTC) were identified in relatively low frequencies.

The REPuter program was used to analyse the nine cp sequences and assess the abundance of four oligonucleotide repeat types:

Table 1

Size comparison of plastome features of nine *Anoectochilus* species.

Accession code	Scientific name	Genome size (bp)	LSC size (bp)	SSC size (bp)	IRB size (bp)	IRB size (bp)	Coding genes	rRNA	tRNA
VN_*A_formosanus*	*A. formosanus*	152 658	82 692	17 346	23 610	23 610	92	10	39
NC_033895.1	*A. emeiensis*	152 650	82 670	17 342	26 319	26 319	93	8	46
NC_066958.1	*A. burmannicus*	152 868	82 733	17 473	26 331	26 331	89	8	38
NC_061758.1	*A. roxburghii*	152 821	82 693	17 488	26 320	26 320	91	8	37
NC_061756.1	*A. formosanus*	151 414	81 879	16 909	26 313	26 313	90	8	37
NC_054353.1	*A. zhejiangensis*	152 509	82 660	17 201	26 324	26 324	90	8	38
MW589501.1	*A. hainanensis*	152 645	82 881	17 626	26 069	26 069	90	8	38
MW589500.1	*A. chapaensis*	152 395	82 630	17 125	26 320	26 320	90	8	38
MT041259.1	*A. calcareus*	151 864	82 083	17 141	26 320	36 320	92	10	39

Table 2

Estimates of evolutionary divergence among cp genome sequences of nine *Anoectochilus* species

No	Accession number	Species	1	2	3	4	5	6	7	8
1	VN_*A_formosanus*	*A. formosanus*								
2	NC_033895.1	*A. emeiensis*	0.000							
3	NC_066958.1	*A. burmannicus*	0.002	0.002						
4	NC_061758.1	*A. roxburghii*	0.002	0.002	0.001					
5	NC_061756.1	*A. formosanus*	0.002	0.002	0.001	0.001				
6	NC_054353.1	*A. zhejiangensis*	0.006	0.005	0.005	0.005	0.005			
7	MW589501.1	*A. hainanensis*	0.005	0.005	0.005	0.005	0.005	0.003		
8	MW589500.1	*A. chapaensis*	0.005	0.005	0.005	0.005	0.005	0.003	0.000	
9	MT041259.1	*A. calcareus*	0.005	0.005	0.004	0.004	0.005	0.005	0.004	0.005

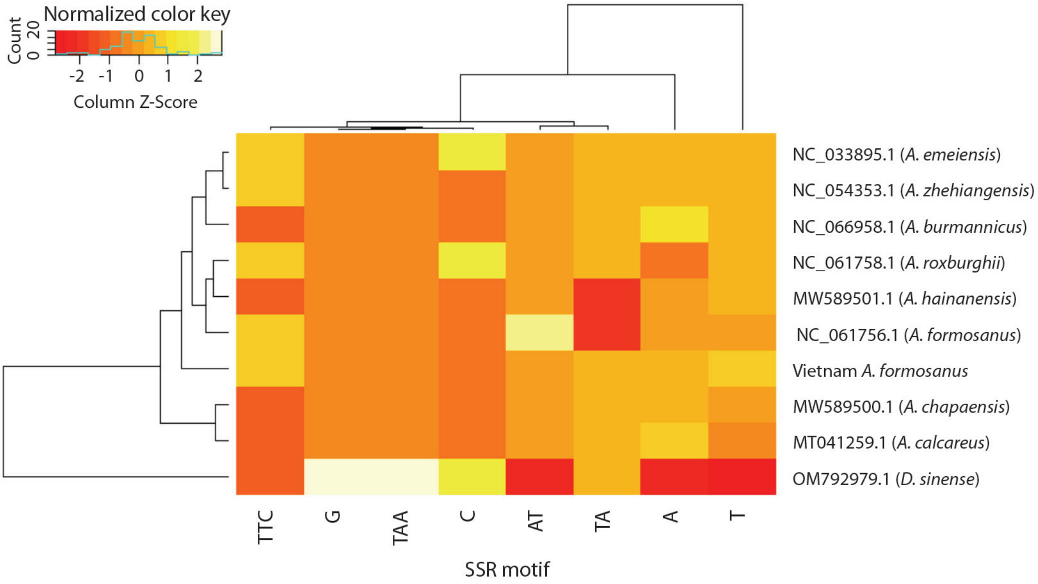

Fig. 2. The different simple sequence repeat types in the cp genomes of nine *Anoectochilus* species and *D. sinense* as an outgroup.

forward (F), palindromic (P), reverse (R), and complementary (C). The number and type of repeat elements exhibited significant variation among the nine cp genomes (Table 3), ranging from 38 units in *A. chapaensis* to 53 units in *A. zhejiangensis*.

IR contraction and expansion: Despite the high conservation of genomic structure and size in the nine *Anoectochilus* cp genomes, the IR/SC boundary regions exhibited notable differences (Fig. 3). Several genes, including *rpl*22, *ndh*F, and *ycf*1, varied in length. Intriguingly, *ndh*F was absent in the IRb/SSC border in only the *A. hainanensis* cp genome, indicating that the loss of this gene likely occurred independently among jewel orchid species in the *Anoectochilus* genus. Although the *rpl*22 gene was present in all LSC/IRb borders, this gene was observed only in 4/9 cp genomes in the IRa region, including *A. burmannicus*, *A. hainanensis*, *A. chapaensis*, and *A. zhejiangensis*.

Table 3
Number of repeated sequences in nine *Anoectochilus* cp genomes.

No	Accession	Forward vs. Forward	Forward vs. Complement	Forward vs. Reverse	Forward vs. Reverse Complement
1	VN_*A_formosanus*	8	7	13	22
2	NC_033895.1	17	0	24	9
3	NC_066958.1	17	2	19	12
4	NC_061758.1	12	5	14	18
5	NC_061756.1	10	7	13	20
6	NC_054353.1	10	3	23	17
7	MW589501.1	18	0	19	12
8	MW589500.1	17	1	10	10
9	MT041259.1	21	2	7	20

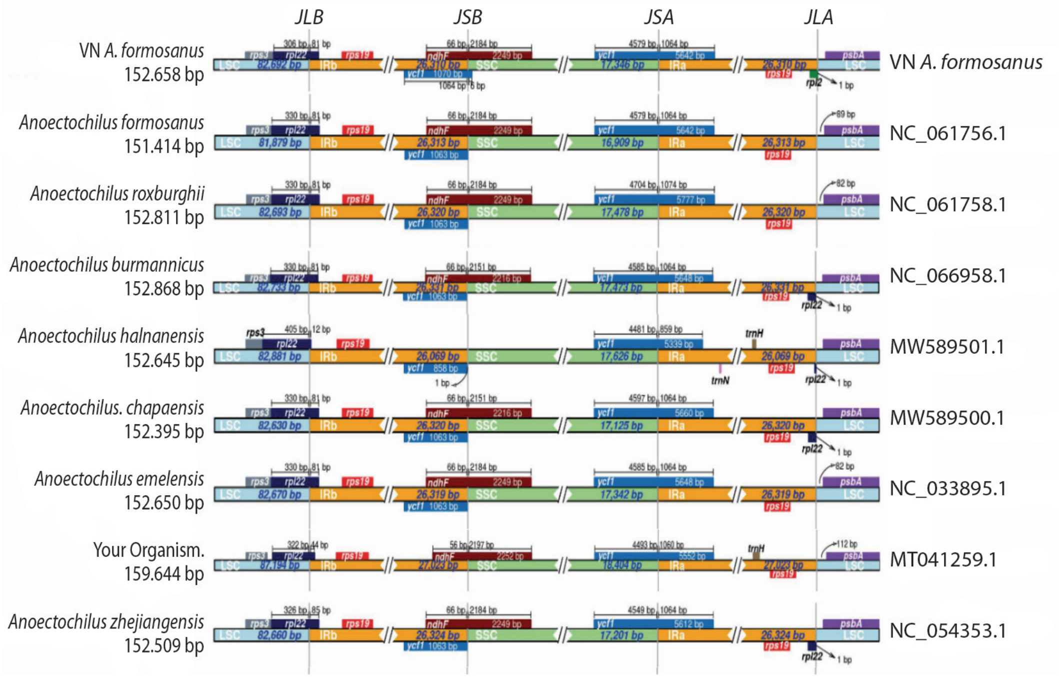

Fig. 3. The LSC, IR, and SSC border regions were compared among the nine *Anoectochilus* cp genomes. Genes located at the IR/SC borders are represented by boxes above or below the main lines, with the numbers above the gene indicating the distance in bp from the gene terminal to the boundary region.

We also noted that *rpl2* gene is only appeared in *A. formosanus* from Vietnam.

In order to align the cp genome of the nine *Anoectochilus* species, the NC_061756.1 cp genome was employed as a reference in sequence alignment with mMISTA software (Fig. 4). Overall, the size and gene order of the nine cp genomes analysed were found to be conserved.

Phylogenetic relationship: The phylogenetic analysis of the nine *Anoectochilus* cp genomes revealed a distinct clustering pattern among jewel orchid species. The cp genomes are divided into two main clades with high bootstrap values (Fig. 5). In both phylogenetic trees built by Neighbor- Joining phylogeny (Fig. 5A) and Maximum Likelihood method (Fig. 5B), *A. formosanus* accession from Vietnam and *A. formosanus* accession from China (NC_061756.1) are formed a sub clade which are sisters with another clade forming by three accessions namely *A. emeiensis* (NC_033895.1), *A. roxburghii* (NC_061758.1) and *A. zhejiangensis* (NC_054353.1). Nevertheless, the is slight difference between two phylogenies where *A. hainanensis* (MW589501.1) is grouped with other three accessions, namely *A. burmannicus* (NC_066958.1), *A. chapaensis* (MW589500.1) and *A. calcareus* (MT041259.1) with the high bootstrap value at 99 to form the second main clade using NJ method (Fig. 5A). Whereas, by using ML method *A. hainanensis* (MW589501.1) is clustered in different clade which are closer to two *A. formosanus* accessions (Fig. 5B).

DISCUSSION

Genetic variation in cp genomes: Orchidaceae is a large family of flowering plants that comprises numerous endangered, rare, and threatened species, including *A. formosanus*. Due to human excavation and habitat

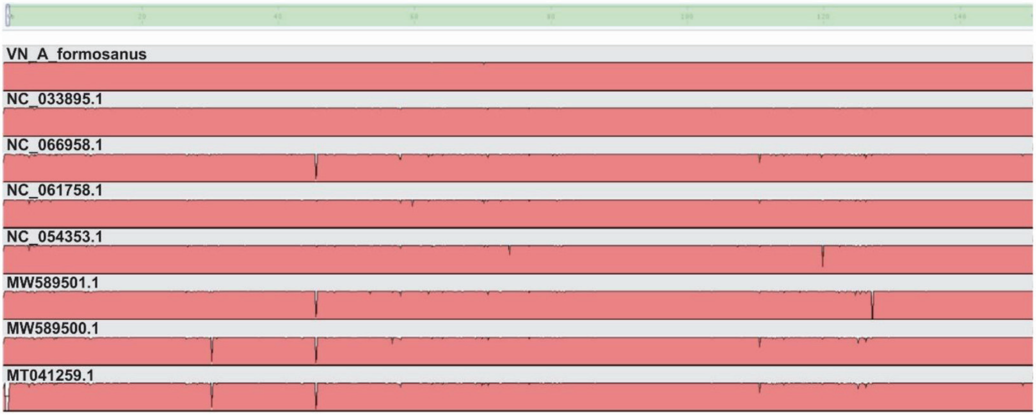

Fig. 4. Sequence identity plot compared nine cp genomes with NC_061756.1 as a reference by using mVISTA. A cut-off of 70 % identity was used for the plots, and the Y-scale represents the percent identity from 50 to 100 %.

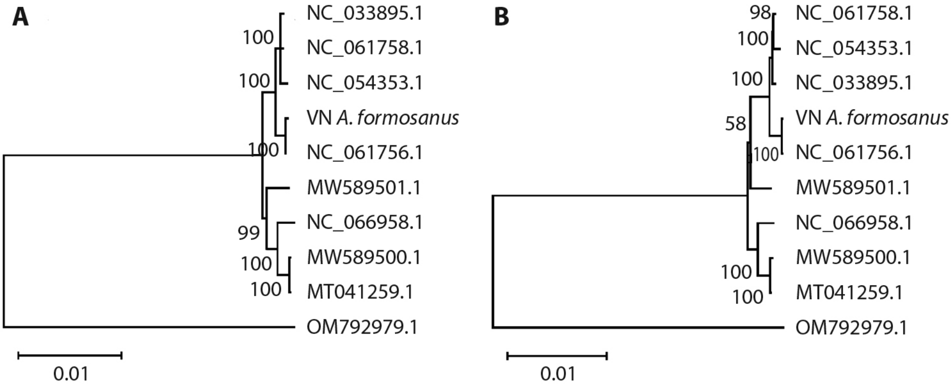

Fig. 5. Phylogenetic tree of nine *Anoectochilus* jewel orchid species using Neighbor-Joining method (**A**) and using the Maximum Likelihood method (**B**). The cp sequences *D. sinense* was used as outgroup. Numbers near branches are bootstrap values.

destruction, the presence of this species is highly vulnerable. Gaining a deeper understanding of the genetic composition of this plant could greatly contribute to improved management and conservation programs. Recently, chloroplast genomes have provided valuable information for understanding genetic diversity, phylogenetics, and speciation in land plants (Henriquez et al., 2022; Liu et al., 2022; Ren et al., 2021). In this study, we sequenced and annotated the complete cp genome of *A. formosanus* for the first time in Vietnam. The assembled plastome exhibits the typical structure of a plant cp genome, consisting of four parts. The GC content, an important parameter for plant identification as higher GC content aids in protecting the cp genome structure, was found to be similar to that of other species in the *Anoectochilus* genus in our study (Nguyen et al., 2023a).

We found the variations among the cp genomes of *A. formosanus* collected from Vietnam and the same species from China (accession: NC_061756.1) in terms of genome size, gene numbers, and sequence repeat motifs. This information is not surprised since previous studies have reported notable variations among accessions within species (Ren et al., 2021;

Xu et al., 2022). The nucleotide diversity value (Pi) in this study was relatively low compared to other species in the orchid family such as *Paphiopedilum* (Liu et al., 2022). The low levels of nucleotide diversity could be due to humans' selection pressure on the plant (Kanaka et al., 2023). The obtained data provides insights into the typical structure and content of the cp genomes of *A. formosanus* in Vietnam. These differences among cp genomes contribute to our understanding of the genetic structure within the *Anoectochilus* genus.

Early studies have reported the abundance of repeated motifs in cp genomes, which are associated with various types of genome rearrangements, recombination, and large inversions, making them valuable for phylogenetic studies (Ho et al. 2023; Lei et al. 2016; Zhang et al., 2022). In our study, we observed variable numbers of repeats among the cp genomes, consistent with findings in a study on 13 Aroideae species where repeat numbers were found to be unrelated to genome size and phylogenetic position of the species (Henriquez et al. 2022). The A and T motifs were commonly identified in the nine *Anoectochilus* cp genomes, aligning with a previous study by Nguyen et al. (2023a). Microsatellites present in the cp genome are inherited from a single parent and are frequently utilized as molecular markers in evolutionary studies, such as assessing genetic diversity and species identification. The new and specific microsatellites from *A. formosanus* accession with Vietnam origin detected in our study hold promise for evolutionary investigations in the *A. formosanus* genus, as well as aiding in the identification and conservation of different species within this genus. These cp SSRs (simple sequence repeats) are informative molecular markers for evaluating genetic relationships due to their high polymorphism and copy number variation. The identified cp SSR loci could prove highly valuable for genetic diversity studies and may enhance the effectiveness of interspecific discrimination, potentially in combination with other nuclear genomic SSRs (Zarei et al., 2022).

In general, cp genomes exhibit high conservation in terms of gene content and organization. However, variations are often observed in the IR regions, indicating the involvement of contraction and expansion events in shaping cp genomes. Our data show a large variation in the presence and absence of *ndh*F and *rpl22* genes in the SSC/IR junctions. The loss of the *ndh*F gene was also reported in other orchid cp genomes (Lin et al., 2017) and other plants such as *Taxillus* (Li et al., 2017), *Buchnera americana* (Frailey et al., 2018). This loss could be attributed to plant evolution towards photosynthetic adaptation (Scobeyeva et al., 2021). On the other hand, previous studies reported that *rpl*22 is one of the genes with the highest deletion rate in plant cp genomes. Daniell et al. (2016) reported the loss of this gene in 57 cp genomes of 26 plant genera, and up to 127 deletions were detected after comparing 2 511 cp genomes (Mohanta et al., 2020). Interestingly, our study found that *rpl*2 gene was absent in China *A. formosanus* (NC_061756.1) which would facilitate the development of molecular markers to differentiate between these two *A. formosanus* accessions.

Phylogenetic relationship: Previous attempts to determine phylogenetic relationships using a limited set of genetic markers were insufficient in accurately establishing these relationships, especially when dealing with closely related species. Specifically, when applying DNA barcoding techniques, misclassification of species in *Anoectochilus* genus has been frequent (Gao et al., 2009). Despite incorporating multiple DNA barcode regions, including *rbc*L, *mat*K, *rpo*B1, *rpo*B2, *rpo*C1, *rpo*C2, ITS1, ITS2, and ITS, Huynh et al. (2019) faced challenges in effectively distinguishing between *A. formosanus* and *A. roxburghii* which is often misidentified as *A. formosanus* in traditional medicine (Ye et al., 2017). The limited discriminatory power of DNA barcode regions can be attributed to the minimal variation observed among these sequences across species, with only a few single nucleotide polymorphisms detected (Chen & Shiau, 2015;

Huynh et al., 2019). The phylogenetic analysis in this study reveals that the two samples of *A. formosanus* formed a distinct subclade separate from *A. roxburghii*. This finding aligns with a previous study by Nguyen et al. (2023a), which also classified *A. formosanus* and *A. roxburghii* into different clades. Consequently, the analysis of complete cp genomes provides a suitable approach for resolving contentious phylogenetic relationships among different species, even at lower taxonomic levels.

In this study we sequenced and characterized the complete cp sequence of *A. formosanus*, an endangered jewel orchid species collected from Vietnam. Comparative analysis of this cp genome with closely related species revealed distinct features, including variations in genome size, gene numbers, and sequence repeat motifs. The obtained data provides valuable insights into the typical structure and content of cp genomes in *A. formosanus* from Vietnam. These differences among cp genomes contribute to our understanding of the genetic structure within the *Anoectochilus* genus. Furthermore, the identification of unique repeat motifs and highly divergent regions in the cp genome of Vietnam *A. formosanus* holds potential for developing molecular markers. These markers can be utilized in future studies focusing on taxonomy and conservation efforts for this precious herb in Vietnam.

Ethical statement: the authors declare that they all agree with this publication and made significant contributions; that there is no conflict of interest of any kind; and that we followed all pertinent ethical and legal procedures and requirements. All financial sources are fully and clearly stated in the acknowledgments section. A signed document has been filed in the journal archives.

ACKNOWLEDGMENT

This work was supported by the Ho Chi Minh City University of Industry and Trade-Vietnam (Formerly Ho Chi Minh City University of Food Industry) through the HUFI fund for Science and Technology under the Contract No. 157/HD-DCT.

REFERENCES

Amiryousefi, A., Hyvönen, J., & Poczai, P. (2018). IRscope: an online program to visualize the junction sites of chloroplast genomes. *Bioinformatics, 34(17),* 3030–3031.

Bhattacharjee, A. B., & Chowdhery, H. J. (2013). Two frequently confused species of 'Jewel orchid' (Orchidaceae - Goodyerinae) from India. *Taiwania, 58*(3), 213–216.

Beier, S., Thiel, T., Münch, T., Scholz, U., & Mascher, M. (2017). MISA-web: A web server for microsatellite prediction. *Bioinformatics, 33*(16), 2583–2585.

Chen, J. R., & Shiau, Y. J. (2015). Application of internal transcribed spacers and maturase K markers for identifying *Anoectochilus*, *Ludisia*, and *Ludochilus*. *Plant Biology, 59*(3), 485–490.

Chiang, S. H., & Lin, C. C. (2017). Antioxidant properties of different portions of organic *Anoectochilus formosanus* Hayata with different drying treatments. *Bioscience Journal- Uberlândia, 34*(1), 12–23.

Daniell, H., Lin, C. S., Yu, M., & Chang, W. J. (2016). Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biology*, *17*(2016), 134.

Frailey, D. C., Chaluvadi, S. R., Vaughn, J. N., Coatney, C. G., & Bennetzen. J. L. (2018). Gene loss and genome rearrangement in the plastids of five *Hemiparasites* in the family Orobanchaceae. *BMC Plant Biology*, *18*, 30.

Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., & Dubchak, I. (2004). VISTA: computational tools for comparative genomics. *Nucleic Acids Research; 32*, W273–W279.

Gao, C., Zhang, F., Zhang, J., Guo, S., & Shao, H. (2009). Identification of *Anoectochilus* based on rDNA ITS sequences alignment. *International Journal of Biological Sciences, 5*(7), 727–75.

Henriquez, C. L., Abdullah, Ahmed, I., Carlsen, M. M., Zuluaga, A., Croat, T. B., & McKain, M. R. (2022). Evolutionary dynamics of chloroplast genomes in subfamily Aroideae (Araceae). *Genomics, 112*(3), 2349–2360.

Ho, Y., Chen, Y. F., Wang, L. H., Hsu, K. Y., Chin, Y. T., Yang, Y. C. S. H., Wang, S. H., Chen, Y. R., Shih, Y. J., Liu, L. F., Wang, K., Whang-Peng, J., Tang, H. Y., Lin, H. Y., Liu, H. L., & Lin, S. J. (2018). Inhibitory effect of *Anoectochilus formosanus* extract on hyperglycemia-related PD-L1expression and cancer proliferation. *Frontiers in Pharmacology*, *9*, 807.

Ho, V. T., Tran, T. K. P., Vu, T. T. T., & Widiarsih, S. (2021). Comparison of *mat*K and *rbc*L DNA barcodes for genetic classification of jewel orchid accessions in Vietnam. *Journal of Genetic Engineering and Biotechnology*, *19*(1), 93.

Ho, V. T., Nguyen, M. P., & Trinh, T. H. (2023). The initial complete chloroplast genome of *Ludisia discolor* (Orchidaceae) in Vietnam. *Nusantara Bioscience, 15*(2), 232–237.

Hu, C., Tian, H., Li, H., Hu, A., Xing, F., Bhattacharjee, A., Hsu, T., Kumar, P., & Chung, S. (2016) Phylogenetic analysis of a 'Jewel Orchid' genus *Goodyera* (Orchidaceae) based on DNA sequence data from nuclear and plastid regions. *PLoS ONE, 11*(2), e0150366

Huynh, H. D., Nguyen, T. G., Duong, H.X., Ha, T. L., Phan. D. Y., Tran, T. T., & Do, D. G. (2019). Using some DNA barcode for the genetic analysis and identifying some species of *Anoectochilus* spp. *Can Tho University Journal of Science, 55*(1), 14–23.

Jiang, J. H., Lee, Y. I., Cubeta, M. A., & Chen, L. C. (2015). Characterization and colonization of endomycorrhizal *Rhizoctonia* fungi in the medicinal herb *Anoectochilus formosanus* (Orchidaceae). *Mycorrhiza, 25*, 431–445.

Kanaka, K. K., Sukhija, N., Goli, R. C., Singh, S., Ganguly, I., Dixit, S. P., Dash, A., & Malik, A. A. (2023). On the concepts and measures of diversity in the genomics are. *Current Plant Biology, 33,* 100278.

Kang, Y., Deng, Z., Zang, R., & Long, W. (2017). DNA barcoding analysis and phylogenetic relationships of tree species in tropical cloud forests. *Scientific Reports*, *7*(2017), 12564.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4),772–80.

Katoh, K., Rozewicki, J., & Yamada, K. D. (2019). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics, 20*(4), 1160–1166.

Konhar, R., Debnath, M., Vishwakarma, S., Bhattacharjee, A., Sundar, D., Tandon, P., Dash, D., & Biswal, D. K. (2019). The complete chloroplast genome of *Dendrobium nobile*, an endangered medicinal orchid from north-east India and its comparison with related *Dendrobium* species. *PeerJ,* 7, e7756.

Kumar, P., & Gale, S. W. (2020). *Anoectochilus formosanus* (Orchidaceae), a new record for Hong Kong. *Journal of the Indian Association for Angiosperm Taxonomy, 30*(2), 293–298.

Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J., & Giegerich, R. (2001). REPuter: The manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Research, 29*(22), 4633–4642.

Lei, W., Ni, D., Wang, Y., Shao, J., Wang, X., Yang, D., Wang, J., Chen, H., & Liu, C. (2016). Intraspecific and heteroplasmic variations, gene losses and inversions in the chloroplast genome of *Astragalus membranaceus*. *Scientific Reports, 22*(6), 21669.

Li, Y., Zhou, J., Chen, X., Cui, Y., Xu, Z., Li, Y., Song, J., Duan, B., & Yao, H. (2017). Gene losses and partial deletion of small single-copy regions of the chloroplast genomes of two hemiparasitic *Taxillus* species. *Scientific Reports, 7*, 12834.

Li, K., Li, Y., & Nakamura, F. (2023). Identification and partial characterization of new cell density-dependent nucleocytoplasmic shuttling proteins and open chromatin. *Scientific Reports, 13*, 21723.

Lin, J. M., Lin, C. C., Chiu, H. F., Yang, J. J., & Lee, S. G. (1993). Evaluation of the anti-inflammatory and liver-protective effects of *Anoectochilus formosanus*, *Ganoderma lucidum* and *Gynostemma pentaphyllum* in rats. *American Journal of Chinese Medicine*, *21*(1), 59–69.

Lin, S. F., Tsay, H. S., Chou, T. W., Yang, M. J., & Cheng, K. T. (2007). Genetic variation of *Anoectochilus formosanus* revealed by ISSR and AFLP analysis. *Journal of Food and Drug Analysis*, *15*(2), 156–162.

Lin, C. S., Chen, J. J. W., Chiu, C. C., Hsiao, H. C. W., Yang, C. J., Jin, X. H., Leebens-Mack, J., de Pamphilis, C. W., Huang, Y. T., Yang, L. H., Chang, W. J., Kui, L., Wong, G. K. S., Hu, J. M., Wang, W., & Shih, M. C. (2017). Concomitant loss of NDH complex-related genes within chloroplast and nuclear genomes in some orchids. *The Plant Journal, 90*(5), 994–1006.

Liu, H., Ye, H., Zhang, N., Ma, J., Wang, J., Hu, G., Li, M., & Zhao, P. (2022). Comparative analyses of chloroplast genomes provide comprehensive insights into the adaptive evolution of *Paphiopedilum* (Orchidaceae). *Horticulturae, 8*(5), 391.

Ma, Z., Li, S., Zhang, M., Jiang, S., & Xiao, Y. (2010). Light intensity affects growth, photosynthetic capability, and total flavonoid accumulation of *Anoectochilus* Plants. *Hort science, 45*(6), 863–867.

Mohanta, T. K., Mishra, A. K., Khan, A., Hashem, A., Abd-Allah, E. F., & Al-Harrasi, A. (2020). Gene loss and evolution of the plastome. *Genes, 11*(10), 1133.

NCBI (National Center for Biotechnology Information). (1982). *GenBank*. https://www.ncbi.nlm.nih.gov/nucleotide/

NCBI (National Center for Biotechnology Information). (2009). *Sequence Read Archive (SRA)*. https://www.ncbi.nlm.nih.gov/sra/

Nguyen, M. P., Trinh, T. H., Ngo, T. K. A., Widiarsih, S., & Ho, V. T. (2023a). *In silico* comparative analysis of the complete chloroplast genome sequences in different jewel orchid species. *Nusantara Bioscience, 15*(1), 12–21.

Nguyen, T. P., Phan, H. N., Do, T. D., Do, G. D., Ngo, L. H., Do, H. D. K., & Nguyen, K. T. (2023b). Polysaccharide and ethanol extracts of *Anoectochilus formosanus* Hayata: Antioxidant, wound-healing, antibacterial, and cytotoxic activities. *Heliyon, 9*(3), e13559.

Nishimaki, T., & Sato, K. (2019). An extension of the Kimura two-parameter model to the natural evolutionary process. *Journal of Molecular Evolution, 87*(2019), 60–67.

Ogier, A., & Dorval, T., (2012). HCS-Analyzer: Open source software for high-content screening data correction and analysis. *Bioinformatics, 28*(14), 1945–1946.

Ong, B., & Lee, C. T. (2019). Diversity and conservation of jewel orchids (*Anoectochilus, Goodyera, Ludisia, and Macodes*) in Peninsular Malaysia. *Journal of Tropical Forest Science, 31*(3), 280–292.

Ren, F., Wang, L., Li, Y., Zhou, W., Xu, Z., Gou, H., Liu, Y., Gao, R., & Song, J. (2021). Highly variable chloroplast genome from two endangered Papaveraceae lithophytes *Corydalis tomentella* and *Corydalis Saxicola*. *Ecology and Evolution, 11*(9), 4158–4171.

Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., & Sánchez-Gracia, A. (2017). DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Molecular Biology and Evolution, 34(12)*, 3299–3302.

Scobeyeva, V. A., Artyushin, I. V., Krinitsina A. A., Nikitin, P. A., Antipin, M. A., Kuptsov, S. V., Belenikin, M. S., Omelchenko, D. O., Logacheva, M. D., Konorov E. A., Samoilov A. E., & Speranskaya A. S. (2021). Gene loss, pseudogenization in plastomes of genus *Allium* (Amaryllidaceae), and putative selection for adaptation to environmental conditions. *Frontiers in Genetics*, *12*, 674783.

Shiau, Y. J., Sarage, A. P., Chen, U. C., Yang, S. R., & Tsay H. S. (2002). Conservation of *Anoectochilus formosanus* Hayata by artificial cross-pollination and in vitro culture of seeds. *Botanical Bulletin of Academia Sinica, 43*(2), 123–130.

Shih, C. C., Wu, Y. W., & Linh, W. C. (2002). Antihyperglycaemic and anti-oxidant properties of *Anoectochilus formosanus* in diabetic rats. *Clinical and Experimental Pharmacology and Physiology, 29*(8), 684–688.

Suetsugu, K., Hirota, S. K., Nakato, N., Suyama, Y., & Serizawa S. (2022). Morphological, ecological, and molecular phylogenetic approaches reveal species boundaries and evolutionary history of *Goodyera crassifolia* (Orchidaceae, Orchidoideae) and its closely related taxa. *PhytoKeys, 212*, 111–134.

Tang, T., Duan, X., Ke, Y., Zhang, L., Shen, Y., Hu, B., Liu, A., Chen, H., Li, C., Wu, W., Shen, L., & Liu, Y. (2018). Antidiabetic activities of polysaccharides from *Anoectochilus roxburghii* and *Anoectochilus formosanus* in STZ-induced diabetic mice. International *Journal of Biological Macromolecules, 112*, 882–888.

Tang, H., Tang, L., Shao, S., Peng, Y., Li, L., & Lue, Y. (2021). Chloroplast genomic diversity in *Bulbophyllum* section *Macrocaulia* (Orchidaceae, Epidendroideae, Malaxideae): Insights into species divergence and adaptive evolution. *Plant Diversity*, *43*(5), 350–361.

The Galaxy Community. (2022). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Research, 50*(W1), W345–W351.

Tran, T. K. P., Pham, M. H., Trinh, T. H., Widiarsih, S., & Ho, V. T. (2022). Investigation of the genetic diversity of jewel orchid in Vietnam using RAPD and ISSR markers. *Biodiversitas, 23*(9), 4816–4825.

Wang, S. Y., Kuo, Y. H., Chang, H. N., Khang, P. L., Tsay, H. S., Lin, K. F., Yang, N. S., & Shyur, L. F. (2002). Profiling and characterization antioxidant activities in *Anoectochilus formosanus* Hayata. *Journal of Agricultural and Food Chemistry, 50*(7), 1859–1865.

Xu, Y., Wen, J., Su, X., & Ren, Z. (2022). Variation among the complete chloroplast genomes of the sumac species *Rhus chinensis*: Reannotation and comparative analysis. *Genes, 13*(11), 1936.

Yang, J. B., Tang, M., Li, H. T., Zhang, Z. R., & Li, D. Z. (2013). Complete chloroplast genome of the genus *Cymbidium*: lights into the species identification, phylogenetic implications and population genetic analyses. *BMC Evolutionary Biology*, *13*, 84.

Ye, S., Shao, Q., & Zhang, A. (2017). *Anoectochilus roxburghii*: A review of its phytochemistry, pharmacology, and clinical applications. *Journal of Ethnopharmacology, 209*, 184–202.

Zarei, A., Ebrahimi, A., Mathur, S., & Lawson, S. (2022). The first complete chloroplast genome sequence and phylogenetic analysis of pistachio (*Pistacia vera*). *Diversity, 14*(7), 577.

Zhang, X., Liu, Y., Chen, Y., & Liu, Z. (2019). A comparison of DNA barcoding and morphological identification of jewel orchids (*Anoectochilus*, *Goodyera* and *Ludisia*). *Plant Diversity, 41*(5), 283–290.

Zhang, F. S., Lv, Y. L., Zhao, Y., & Guo, S. X. (2013). Promoting role of an endophyte on the growth and contents of kinsenosides and flavonoids of *Anoectochilus formosanus* Hayata, a rare and threatened medicinal Orchidaceae plant. *Journal of Zhejiang*

University-SCIENCE B (Biomedicine & Biotechnology), 14(9), 785–792.

Zhang, J. Y., Liao, M., Cheng, Y. H., Feng, Y., Ju, W. B., Deng, H. N., Li, X., Plenkovic-Moraj, A., & Xu, B. (2022). Comparative chloroplast genomics of seven endangered cypripedium species and phylogenetic relationships of Orchidaceae. *Frontiers in Plant Science,* 13, 911702.

Zheng, S., Poczai, P., Hyvönen, J., Tang, J., & Amiryousefi, A. (2020). Chloroplot: An online program for the versatile plotting of organelle genomes. *Frontiers in Genetics, 11*, 576124.