EDINBURGH JOURNAL OF BOTANY 79, *Begonia* special issue,
Article 408: 1–32 (2022). https://doi.org/10.24823/EJB.2022.408

Published by the Royal Botanic Garden Edinburgh
ISSN (online): 1474-0036, ISSN (print): 0960-4286

OPEN ACCESS
PEER REVIEWED

INSIGHTS INTO THE EVOLUTION OF THE CHLOROPLAST GENOME AND THE PHYLOGENY OF *BEGONIA*

Y.-H. Tseng[1,2], C. L. Hsieh[1], L. Campos-Domínguez[3–5], A. Q. Hu[1,6], C. C. Chang[1], Y. T. Hsu[7], C. A. Kidner[3,4], M. Hughes[3], P. W. Moonlight[3], C. H. Hung[7], Y. C. Wang[8], Y. T. Wang[9], S. H. Liu[10], D. Girmansyah[11] & K.-F. Chung[1*]

Begonia (Begoniaceae) is one of the largest angiosperm genera, comprising more than 2000 species; this makes it ideal as a model to investigate the genomic basis of species radiations. Here we present the results of the first genus-wide comparative study of plastid genome structure, sequence diversity, and phylogenetics of Begoniaceae, in which 44 complete Begoniaceae plastomes, including those of *Begonia*'s sister group, *Hillebrandia*, a monotypic genus endemic to Hawaiʻi, and 43 species representing 42 sections of *Begonia*, were assembled. Our results reveal that Begoniaceae plastome size ranges from 167,123 to 170,852 bp, displaying the typical quadripartite structure. Structures of most Begoniaceae plastomes are highly conserved but differ from the plastomes of the majority of angiosperms in having a unique inverted repeat (IR) expansion, from IRa to large single copy (LSC), resulting from a duplicated fragment of the *trnH–GUG* gene to the *trnR–UCU* gene. Additionally, comparison between plastomes of *Hillebrandia* and *Begonia* shows that the former genus has fewer simple sequence repeats than most *Begonia* species analysed, suggesting that species of *Begonia* have more repetitive and dynamic plastomes than those of its sister genus. We also identified six highly variable regions suitable for phylogenetic analysis and as potential DNA barcodes for species identification. Our robust hypothesis of plastome phylogenomic relationships provides new insights into infrageneric classification and highlights potential classification issues in *Begonia*.

Keywords. *Begonia*, Begoniaceae, *Hillebrandia*, IR expansion, plastome, species barcode.
Received 10 August 2021 Accepted 20 January 2022 Published 18 August 2022

[1] Research Museum and Herbarium (HAST), Biodiversity Research Center, Academia Sinica, 128 Academia Road, Section 2, Taipei 115201, Taiwan.
[2] Department of Life Sciences, National Chung Hsing University, 145 Xingda Road, South District, Taichung City 402202, Taiwan.
[3] Royal Botanic Garden Edinburgh, 20A Inverleith Row, Edinburgh EH3 5LR, Scotland, UK.
[4] Institute of Molecular Plant Sciences, University of Edinburgh, The King's Buildings, Edinburgh EH9 3BF, Scotland, UK.
[5] Institute of Evolutionary Biology, School of Biological Sciences, Ashworth Laboratories, University of Edinburgh, Edinburgh EH9 3FL, Scotland, UK.
[6] Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3AE, England, UK.
[7] Department of Life Sciences, National Cheng Kung University, 1 University Road, Tainan 701401, Taiwan.
[8] School of Forestry and Resource Conservation, National Taiwan University, 1, Section 4, Roosevelt Road, Taipei 106319, Taiwan.

Introduction

The chloroplast is an essential organelle in plants, carrying out photosynthesis as well as biosynthesis of fatty acids and starch (Gray, 1989; Kleffmann *et al.*, 2004). Angiosperm plastid genomes (plastomes) are typically circular, are maternally inherited, and have a quadripartite structure. They typically range from 130 to 170 kb in length, and are organised into one large single copy (LSC) and one small single copy (SSC) region, flanked by two inverted repeats (IRa and IRb) (Palmer, 1987; Green, 2011). The plastomes of angiosperms are mostly conserved in terms of gene content and structure, and yet there is considerable variation resulting from the expansion and contraction of IRs (Sun *et al.*, 2013), the addition and deletion of genes (McNeal *et al.*, 2007), the inversion of genes and regions (Park *et al.*, 2018), and polymorphic simple sequence repeats (SSRs) (Cheng *et al.*, 2016).

The biological characteristics of the plastome, including its uniparental inheritance, the absence of recombination, and its low rate of nucleotide substitution, make it ideal for ecological and evolutionary studies (Twyford & Ness, 2017). The number of fully sequenced and annotated plastomes is rapidly increasing, and these data are being used in plant phylogenetics to address evolutionary questions across various ranks and taxa. So far, more than 2700 full plastid genome sequences have been published (Asaf *et al.*, 2020).

Plastomes in Begoniaceae

Begonia L. (Begoniaceae) is one of the largest angiosperm genera, and comprises more than 2000 species divided among 70 sections (Hughes *et al.*, 2015–; Moonlight *et al.*, 2018). It is distributed worldwide in tropical and subtropical Asia, Africa and the Americas (Tebbitt, 2005; Dewitte *et al.*, 2011). The high species diversity of *Begonia* is in contrast to its sister genus, *Hillebrandia* Oliv., which is monotypic (*Hillebrandia sandwicensis* Oliv.) and the only taxon of the Begoniaceae native to the Hawaiian Islands (Clement *et al.*, 2004). *Begonia* exhibits a large range of morphological diversity, particularly with respect to leaf shape, colour and variegation. As a megadiverse, pantropically distributed genus, *Begonia* provides an excellent system for investigating the processes and patterns underlying the generation of biodiversity. To gain insight into the potential role that the plastid genome plays in species differences, a comparative analysis of plastome structure and repeat content was required.

[9] Department of Horticulture, National Chung Hsing University, 145 Xingda Road, Taichung 402202, Taiwan.

[10] Department of Biological Sciences, National Sun-Yat-sen University, 70 Lienhai Road, Kaohsiung 804201, Taiwan.

[11] Research Center for Biosystematics and Evolution, Research Organization for Life Science and Environment, National Research and Innovation Agency (BRIN), Cibinong Science Center, Jalan Raya Jakarta Bogor, Km 46, Cibinong, West Java, 16911, Indonesia.

* Corresponding author. E-mail: bochung@gate.sinica.edu.tw.

In addition to the knowledge of plastome structure and variation, a stable and natural infrageneric classification is required as a basis for studies investigating the factors influencing speciation in *Begonia* (Moonlight *et al.*, 2018). However, *Begonia* taxonomy is remarkably challenging because of the large number of species and the poor preservation of morphological features in specimens (Hughes & Girmansyah, 2011; Chung *et al.*, 2014). To date, extensive phylogenetic studies of *Begonia* and the Begoniaceae have been carried out, examining biogeography, species delimitation, sectional assignment and population genetics. Such studies are generally based on plastid sequences, because nuclear DNA phylogenies provide insufficient resolution, due to substitution saturation across the genus and phylogenetic incongruence derived from frequent natural hybridisation (Moonlight *et al.*, 2018).

Earlier works used plastid sequences from the *trnL* intron (Plana, 2003; Plana *et al.*, 2004; in combination with nrITS) and the *rbcL* region (Clement *et al.*, 2004; Goodall-Copestake *et al.*, 2009; in combination with the nrITS and 18S rRNA gene, respectively) to evaluate sectional delimitation, divergence time and biogeographical patterns. The results of a subsequent phylogenetic study of *Begonia* using five plastid sequences (*trnK* intron/*matK* gene, *petB*–*petD* spacer, *psbB* gene, *psbC*–*trnS* spacer and *trnL* intron) with five mitochondrial sequences (*cox1* gene, *matR* gene, *nad1* gene, *nad7* gene and *rps14*–*cob* spacer) suggested that extant *Begonia* lineages first diversified in Africa and that the closest African relatives of the American and Asian *Begonia* are seasonally dry adapted species (Goodall-Copestake *et al.*, 2010).

Later, Thomas *et al.* (2011) successfully amplified three highly variable plastid sequences (*ndhA* intron, *ndhF*–*rpl32* spacer and *rpl32*–*trnL* spacer) to reconstruct the first supported phylogenetic framework for Asian *Begonia*. Moonlight *et al.* (2018) also utilised these three plastid markers, with a more comprehensive taxon sampling including 574 species of *Begonia*, to establish the first sectional classification based on phylogenetic data, in which 70 sections of *Begonia* were recognised. Additionally, more plastid markers have been applied for species-level phylogenetic studies in *Begonia* sect. *Coelocentrum* (*rpl16* intron with nrITS; Chung *et al.*, 2014); hybridisation detection and biogeography in *Begonia* sect. *Baryandra* (*trnC*–*trnD* spacer with *ndhA* intron, *ndhF*–*rpl32* spacer and *rpl32*–*trnL* spacer; Hughes *et al.*, 2015, 2018); identification of the first natural hybrid in *Begonia* sect. *Petermannia* (*trnL*–*trnF* spacer with nrITS; Liu *et al.*, 2019); and population genetics in *Begonia luzhaiensis* T.C.Ku (*trnC*–*ycf6* spacer; Tseng *et al.*, 2019).

In summary, 13 plastid genes or spacer sequences have been utilised in studies of *Begonia* and Begoniaceae. Although these phylogenetic studies have provided sufficient resolution at the sectional level, resolution at the species level in most sections has proven recalcitrant (Harrison *et al.*, 2016).

Harrison *et al.* (2016) first attempted to assemble plastomes of 16 species of *Begonia* using long-range PCR, but only one nearly complete plastome assembly (*B. peltata* Otto & A.Dietr.) was successfully generated. Subsequently, five additional complete plastomes of Asian *Begonia* were reported (Dong *et al.*, 2019; Fan *et al.*, 2019; Huang & Wang, 2020; Zhou *et al.*, 2020; Wang *et al.*, 2021) based on the high-copy fraction of plastome sequences using the NGS genome skimming method (Straub *et al.*, 2012). The sizes of these plastome assemblies ranged from 157,648 to 169,436 bp, and they have a typical quadripartite structure (Dong *et al.*, 2019; Fan *et al.*, 2019; Huang & Wang, 2020; Zhou *et al.*, 2020; Wang *et al.*, 2021). Recently, Shui *et al.* (2019) used 115 taxa covering 98 species of *Begonia* to establish the first plastome phylogeny in the genus, based on which a new infrageneric classification was proposed, although the plastome sequences were not released. However, so far, no comparative study of *Begonia* plastid sequence diversity and structure has been carried out. Furthermore, there has to date been no analysis of plastid sequence diversity to inform the choice of appropriate phylogenetic markers in *Begonia*.

In the present study, we report complete plastomes of *Hillebrandia sandwicensis* and 43 species of *Begonia*, representing 42 of the 70 sections recognised by Moonlight *et al.* (2018). Using these data, we aimed to: (i) characterise and compare plastome structure and gene organisation; (ii) identify putative repeated regions; (iii) identify candidate molecular markers for further phylogenetic analyses; and (iv) reconstruct plastome phylogenomic relationships to improve our understanding of plastome characteristics, structural diversity and evolution in Begoniaceae.

Materials and methods

Taxon sampling

Our samples comprised 44 species of Begoniaceae. We chose 43 *Begonia* species from Asia (12), the Americas (20) and Africa (11), representing 42 of the 70 sections (Hughes *et al.*, 2015–; Moonlight *et al.*, 2018; Krishna *et al.*, 2020) in *Begonia*: sections *Astrothrix*, *Augustia*, *Baccabegonia*, *Baryandra*, *Begonia*, *Bracteibegonia*, *Coelocentrum*, *Cyathocnemis*, *Diploclinium*, *Donaldia*, *Erminea*, *Eupetalum*, *Exalabegonia*, *Flocciferae*, *Gaerdtia*, *Gireoudia*, *Haagea*, *Hydristyles*, *Jackia*, *Knesebeckia*, *Latistigma*, *Lepsia*, *Loasibegonia*, *Nerviplacentaria*, *Parietoplacentalia*, *Parvibegonia*, *Peltaugustia*, *Petermannia*, *Pilderia*, *Platycentrum*, *Pritzelia*, *Reichenheimia*, *Ridleyella*, *Rossmannia*, *Ruizopavonia*, *Scutobegonia*, *Squamibegonia*, *Tetrachia*, *Tetraphila*, *Trachelocarpus*, *Urniformia* and *Wageneria*. *Hillebrandia sandwicensis* was also included in this study. The samples were obtained from living plants cultivated in the experimental greenhouse at the Biodiversity Research Center, Academia Sinica (BRCAS), Taipei, Taiwan, and the Royal Botanic Garden Edinburgh, UK. Species and collection information in this study are summarised in the **Appendix table**.

DNA extraction, library preparation and sequencing

Total genomic DNA was extracted using the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany) with some modifications. The DNA concentration was measured by Qubit

3.0 Fluorometer (Thermo Scientific, Massachusetts, USA) and NanoDrop 2000 Spectrophotometer (Thermo Scientific). Approximately 1–1.5 μg per DNA sample was sheared by Bioruptor UCD-200TM (Cosmo-bio Inc., Tokyo, Japan) into fragments of about 200–300 bp, according to the manufacturer's instructions. Dual-indexed libraries were made using NEBNext Ultra II DNA Library Prep Kit (New England BioLabs, Massachusetts, USA), following the 200- to 300-bp insert size protocol. The mean length was evaluated by Fragment Analyzer 5200 (Agilent, California, USA) and quantified using the Qubit Fluorometer. These libraries were sequenced by Illumina HiSeq System, using 150-bp paired end reads at the NGS High Throughput Genomic Core of BRCAS.

Plastome assembly and annotation

The raw reads were quality checked using FastQC (Andrews, 2010). Trimmomatic version 0.39 (Bolger *et al.*, 2014) was used to remove adapters and filter out low-quality read. Bases with a quality score < 20 were removed from the beginning and end of each read, and a sliding window (size = 4 bp) was used to clip reads once the mean quality was < 20. Only reads > 36 bp in length were retained. *De novo* assembling of the plastome was implemented using GetOrganelle pipeline (Jin *et al.*, 2020) for *Begonia* species and using NOVOplasty (Dierckxsens *et al.*, 2017) for *Hillebrandia sandwicensis*. Subsequently, to verify quality and correct assembly errors, all raw reads were mapped to the complete draft plastome generated by GetOrganelle using 'Map to Reference' with default settings in Geneious Prime version 2019.2.1.

Complete assembled plastomes were annotated using the GeSeq web application (Tillich *et al.*, 2017) and manually checked and adjusted for the start and stop codons of each gene in Geneious Prime, using the plastome sequence of *Begonia peltata* (Harrison *et al.*, 2016) as a reference. The tRNA genes were further checked by referring to the secondary structures drawn by tRNAscan-SE web server (Chan & Lowe, 2019). The boundaries of LSC, SSC, IRa and IRb were manually analysed in Geneious Prime. A graphical representation of each plastome with annotation was created in OGDRAW version 1.3.1 (Greiner *et al.*, 2019).

Comparative plastome and sequence divergence analysis

To investigate IR expansion or contraction, we compared the boundaries between IR and SC regions of the Begoniaceae, Cucurbitaceae [*Gynostemma pentaphyllum* (Thunb.) Makino, NCBI Reference Sequence (RefSeq) accession number: KT695603] and *Arabidopsis thaliana* (L.) Heynh. (NCBI RefSeq: AP000423) in Geneious Prime. Additionally, plastome sequences of *Begonia* were used for the sliding window analysis to evaluate nucleotide sequence diversity (π). Nucleotide ambiguities were removed, and subsequently, sequences were aligned by MAFFT version 7.45 (Katoh & Standley, 2013) and manually adjusted in Mesquite version 3.5 (Maddison & Maddison, 2015). The sliding windows analyses were performed in

DnaSP version 6.10 (Rozas *et al.*, 2017) with step size of 200 bp and window length of 600 bp. The variable and parsimony-informative sites of potential DNA barcode were calculated by AMAS (Borowiec, 2016).

Characterisation of simple sequence repeats and repeat structure

The number and location of SSRs in the plastomes were identified by a MISA perl script (Beier *et al.*, 2017). The minimum repeat sizes were set as 10, 5 and 4 units for mono-, di- and trinucleotide SSRs, respectively, and three units for each tetra-, penta- and hexanucleotide SSR. The size and types of the repeating sequence (forward, reverse, palindromic and complement) were analysed using REPuter with a 30-bp minimum repeat size (defined as long repeat here) and a sequence identity ≤ 90% (Hamming distance = 3) (Kurtz *et al.*, 2001).

Phylogenetic analyses

A total of 44 plastome sequences assembled in this study (excluding one IR) and two outgroups of Cucurbitaceae from NCBI, *Ampelosicyos humblotii* (Cogn.) Jum. & H.Perrier (NCBI RefSeq: MN542396) and *Gynostemma pentaphyllum* (NCBI RefSeq: KT695603), were aligned using MAFFT with default settings and subsequently manually adjusted in Mesquite. Phylogenetic analyses were conducted using maximum likelihood (ML) and Bayesian inference (BI) methods. ML analysis was performed using RAxML version 8.2.4 (Stamatakis, 2014), based on the GTR+GAMMA model with 10 searches for the best tree and 1000 standard bootstrap (BS) replicates. BI analysis was based on Markov chain algorithm implemented in MRBAYES version 3.2.7. (Ronquist *et al.*, 2012). Four chains of the Markov chain Monte Carlo simulation under the GTR+GAMMA model were performed for 10,000,000 generations each, with trees sampled every 1000 generations. Before the node probability was calculated (posterior probability, PP), the first 25% sampled trees were discarded.

Results

Assembly and characteristics of plastomes

The Begoniaceae plastomes ranged in size from 167,123 bp (*Begonia meyeri-johannis* Engl.) to 170,852 bp (*B. dipetala* Graham), with a mean of 169,426 bp, and coverage ranged from 64.4 ± 25.1 (*B. kingiana* Irmsch.) to 1,498.2 ± 296.8 (*B. anisosepala* Hook.f.) (**Table 1**). All assembled plastomes of Begoniaceae displayed the typical quadripartite structure of angiosperms, consisting of LSC ranging from 74,787 bp (*Begonia meyeri-johannis*) to 77,328 bp (*B. dipetala*), SSC from 17,464 bp (*Hillebrandia sandwicensis*) to 18,503 bp (*B. aconitifolia* A.DC.), and a pair of IRs from 37,127 bp (*B. henryi* Hemsl.) to 37,748 bp [*B. convolvulacea* (Klotzsch ex Klotzsch) A.DC.] (**Table 1** and **Figure 1**). There was no difference between the plastome structures of *Begonia* and *Hillebrandia* (see **Figure 1**).

Table 1. Begoniaceae plastome sequences characterised in the present study

Genus and species	Section	Continent(s)	Length (bp)					GC content (%)	Coverage
			Total	LSC	IR	SSC	Coding region		
Begonia									
B. aconitifolia A.DC.	*Latistigma*	Americas	170,032	76,481	37,620	18,503	80,454	35.5	480.4 ± 231.4
B. albococcinea Hook.	*Flocciferae*	Asia	170,261	76,640	37,600	18,421	80,280	35.5	236.7 ± 990.9
B. amoeboides Moonlight	*Cyathocnemis*	Americas	168,978	76,038	37,539	17,862	80,189	35.6	419.4 ± 60.2
B. ampla Hook.f.	*Squamibegonia*	Africa	169,505	75,854	37,696	18,258	74,811	35.2	135.4 ± 831.0
B. anemoniflora Irmsch.	*Eupetalum*	Americas	169,158	76,479	37,372	17,935	80,296	35.4	415.6 ± 152.4
B. anisosepala Hook.f.	*Scutobegonia*	Africa	169,555	76,878	37,368	18,137	80,448	35.6	1,498.2 ± 296.8
B. baccata Hook.f.	*Baccabegonia*	Africa	169,961	76,273	37,711	18,267	79,116	35.2	304.0 ± 22.2
B. bogneri Ziesenh.	*Erminea*	Africa	170,250	76,546	37,659	18,386	80,331	35.4	436.8 ± 54.3
B. bracteata Jack	*Bracteibegonia*	Asia	169,797	76,374	37,587	18,249	80,256	35.5	137.0 ± 23.0
B. buddleiifolia A.DC.	*Pilderia*	Americas	168,997	76,063	37,536	17,862	80,181	35.6	795.7 ± 85.5
B. chlorosticta Sands	*Petermannia*	Asia	170,626	76,964	37,694	18,274	79,968	35.4	435.9 ± 623.2
B. convolvulacea (Klotzsch ex Klotzsch) A.DC.	*Wageneria*	Americas	168,493	74,915	37,748	18,082	80,331	35.5	186.9 ± 30.1
B. cubensis Hassk.	*Begonia*	Americas	169,672	76,374	37,564	18,170	80,316	35.5	634.7 ± 74.5
B. depauperata Schott	*Trachelocarpus*	Americas	169,210	76,299	37,503	17,905	80,400	35.3	131.1 ± 29.5
B. dipetala Graham	*Haagea*	Asia	170,852	77,328	37,596	18,333	80,265	35.5	392.8 ± 52.0
B. dregei Otto & A.Dietr.	*Augustia*	Africa	169,439	76,392	37,596	17,855	80,451	35.6	278.6 ± 41.4
B. egregia N.E.Br.	*Tetrachia*	Americas	168,626	75,437	37,521	18,134	80,426	35.6	833.7 ± 129.3
B. fenicis Merr.	*Baryandra*	Asia	168,696	75,641	37,271	18,495	81,533	35.5	93.8 ± 42.6
B. fissistyla Irmsch.	*Hydristyles*	Americas	169,526	76,379	37,518	18,111	80,163	35.5	303.1 ± 55.0
B. foliosa Kunth	*Lepsia*	Americas	169,263	76,214	37,466	18,117	79,665	35.5	458.6 ± 64.5
B. henrilaportei Scherber. & Duruiss.	*Nerviplacentaria*	Africa	170,355	76,569	37,682	18,422	80,268	35.4	428.2 ± 131.3
B. henryi Hemsl.	*Reichenheimia*	Asia	168,124	75,635	37,127	18,235	80,274	35.7	163.3 ± 162.1
B. heydei C.DC.	*Urniformia*	Americas	170,025	76,795	37,519	18,192	80,349	35.4	957.2 ± 866.6
B. karwinskyana A.DC.	*Gireoudia*	Americas	169,815	76,407	37,608	18,192	80,463	35.4	98.8 ± 14.9
B. kingiana Irmsch.	*Ridleyella*	Asia	170,692	77,073	37,716	18,187	80,304	35.4	64.4 ± 25.1
B. komoensis Irmsch.	*Tetraphila*	Africa	167,956	75,502	37,435	17,584	80,169	35.4	80.9 ± 194.4
B. ludwigii Irmsch.	*Knesebeckia*	Americas	170,355	76,946	37,627	18,152	80,430	35.3	172.8 ± 92.1
B. meyeri-johannis Engl.	*Exalabegonia*	Africa	167,123	74,787	37,167	18,002	80,217	35.7	228.0 ± 144.7
B. microsperma Warb.	*Loasibegonia*	Africa	169,651	76,966	37,395	17,895	80,334	35.6	325.0 ± 37.2
B. myanmarica C.I Peng & Y.D.Kim	*Platycentrum*	Asia	168,595	75,802	37,228	18,337	80,478	35.5	157.5 ± 512.6
B. oaxacana A.DC.	*Parietoplacentalia*	Americas	169,783	76,292	37,608	18,275	80,326	35.4	428.6 ± 311.7
B. oxyloba Welw. ex Hook.f.	*Exalabegonia*	Africa	169,003	76,249	37,238	18,278	80,300	35.5	248.0 ± 82.6
B. picturata Yan Liu, S.M.Ku & C.I Peng	*Coelocentrum*	Asia	169,678	76,464	37,551	18,112	80,442	35.4	96.9 ± 17.7
B. ravenii C.I Peng & Y.K.Chen	*Diploclinium*	Asia	169,055	76,242	37,487	17,839	80,286	35.6	260.6 ± 429.7
B. rossmanniae A.DC.	*Rossmannia*	Americas	169,232	76,352	37,581	17,718	80,223	35.5	163.0 ± 52.6
B. samhaensis M.Hughes & A.G.Mill.	*Peltaugustia*	Africa	169,413	76,204	37,489	18,231	80,241	35.5	492.0 ± 137.1
B. sanguinea Raddi	*Pritzelia*	Americas	168,106	75,482	37,397	17,830	80,382	35.7	177.6 ± 38.0
B. santos-limae Brade	*Astrothrix*	Americas	169,747	77,013	37,238	18,258	80,278	35.3	298.4 ± 61.8
B. tigrina Kiew	*Jackia*	Asia	170,169	76,558	37,663	18,285	80,325	35.2	889.7 ± 2549.1
B. ulmifolia Willd.	*Donaldia*	Americas	168,935	75,476	37,689	18,081	80,304	35.4	140.1 ± 52.8
B. undulata Schott	*Gaerdtia*	Americas	169,710	76,083	37,650	18,327	80,484	35.5	358.8 ± 43
B. variabilis Ridl.	*Parvibegonia*	Asia	169,951	76,396	37,603	18,349	80,310	35.5	393.8 ± 1504.3
B. sp. nov., sect. *Ruizopavonia*	*Ruizopavonia*	Americas	169,507	76,235	37,562	18,151	80,301	35.5	226.9 ± 46.9
Hillebrandia									
H. sandwicensis Oliv.	–	Americas	168,863	76,015	37,692	17,464	80,652	35.8	1224

IR, inverted repeat; LSC, large single copy; SSC, small single copy.

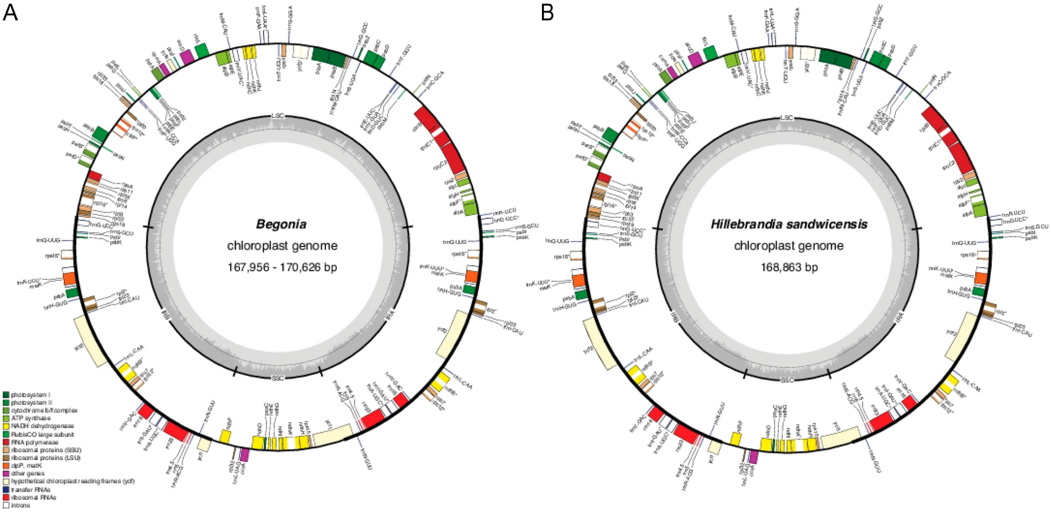


Figure 1. Gene maps for the chloroplast genomes of (A) *Begonia* (represented by *B. fenicis*) and (B) *Hillebrandia sandwicensis*. Genes on the inside and outside of each circle are transcribed in the clockwise and counterclockwise direction, respectively. The dark grey areas within the inner circle indicate the GC content.

The plastomes assembled in this study contained 112 unique genes, including 78 unique protein-coding genes, 30 tRNA genes and four rRNA genes. There were 28 duplicated genes within the IR regions, including 12 protein-coding genes (*matK*, *ndhB*, *psbA*, *psbI*, *psbK*, *rpl2*, *rpl23*, *rps7*, *rps12*, *rps16*, *ycf1* and *ycf2*), 12 tRNAs (*trnA–UGC*, *trnG–UCC*, *trnH–GUG*, *trnI–CAU*, *trnI–GAU*, *trnK–UUU*, *trnL–CAA*, *trnN–GUU*, *trnQ–UUG*, *trnR–ACG*, *trnS–GCU* and *trnV–GAC*), and four rRNA (*rrn4.5*, *rrn5*, *rrn16* and *rrn23*). Among the 112 unique genes in these plastomes, 11 protein-coding genes (*clpP*, *ndhA*, *ndhB*, *petB*, *petD*, *rpl2*, *rpl16*, *rpoC1*, *rps12*, *rps16* and *ycf3*) and six rRNA genes (*trnA–UGC*, *trnG–UCC*, *trnI–GAU*, *trnK–UUU*, *trnL–UAA* and *trnV–UAC*) contained one intron, whereas three genes contained two introns (*cpl*P, *rps12* and *ycf3*). Gene functions and types in the 44 Begoniaceae plastomes are shown in Table 2.

The organisation and IR boundaries in the 44 Begoniaceae plastome sequences were highly conserved, especially in the boundary of IR/SSC. The IRb/SSC boundary was between partial *ycf1* (1224–1460 bp) and *ndhF*. The *ycf1* gene crossed over the IRa/SSC boundary and extended into the IRa region ranging from 1212 to 1416 bp. In most Begoniaceae species, the IRa/LSC boundary was located between *trnG–UCC* and *trnR–UCU* and the IRb/LSC boundary between *trnG–UCC* and *rps19* (represented by *Begonia fenicis* Merr. and *Hillebrandia sandwicensis* in Figure 2). Compared with plastome sequences of *Arabidopsis thaliana* and *Gynostemma pentaphyllum* (see Figure 2), Begoniaceae had an IR expansion, from IRa to LSC including the *trnH–GUG* gene to the *trnG–UCC* gene (*trnH–GUG*, *psbA*, *matK*,

Table 2. Genes identified in 44 Begoniaceae plastomes

Gene function and gene type	Gene name(s)
Photosynthesis	
Photosystem I	*psaA*, *psaB*, *psaC*, *psaI*, *psaJ*, *ycf3*
Photosystem II	*psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbJ*, *psbK*, *psbL*, *psbM*, *psbN*, *psbT*, *psbZ*
Cytochrome b/f complex	*petA*, *petB*, *petD*, *petG*, *petL*, *petN*
ATP synthase	*accD*, *atpA*, *atpB*, *atpE*, *atpF*, *atpH*, *atpI*
NADH dehydrogenase	*ndhA*, *ndhB*, *ndhC*, *ndhD*, *ndhE*, *ndhF*, *ndhG*, *ndhH*, *ndhI*, *ndhJ*, *ndhK*
Rubisco	*rbcL*
Self-replication	
Ribosomal RNA genes	*rrn4.5*, *rrn5*, *rrn16*, *rrn23*
Transfer RNA genes	*trnA–UGC*, *trnC–GCA*, *trnD–GUC*, *trnE–UUC*, *trnF–GAA*, *trnfM–CAU*, *trnG–GCC*, *trnG–UCC*, *trnH–GUG*, *trnI–CAU*, *trnI–GAU*, *trnK–UUU*, *trnL–CAA*, *trnL–UAA*, *trnL–UAG*, *trnM–CAU*, *trnN–GUU*, *trnP–UGG*, *trnQ–UUG*, *trnR–ACG*, *trnR–UCU*, *trnS–GCU*, *trnS–GGA*, *trnS–UGA*, *trnT–GGU*, *trnT–UGU*, *trnV–GAC*, *trnV–UAC*, *trnW–CCA*, *trnY–GUA*
Small ribosomal subunit	*rps2*, *rps3*, *rps4*, *rps7*, *rps8*, *rps11*, *rps12*, *rps14*, *rps15*, *rps16*, *rps18*, *rps19*
Large ribosomal subunit	*rpl2*, *rpl14*, *rpl16*, *rpl20*, *rpl22*, *rpl23*, *rpl32*, *rpl33*, *rpl36*
RNA polymerase subunits	*rpoA*, *rpoB*, *rpoC1*, *rpoC2*
Other genes	
Maturase	*matK*
Protease	*clpP*
Envelope membrane protein	*cemA*
Subunit of acetyl-CoA-carboxlyase	*accD*
Cytochrome c biogenesis protein	*ccsA*
Component of TIC complex	*ycf1*
Unknown function	*ycf2*, *ycf4*

trnK–UUU, *rps16*, *trnQ–UUG*, *psbK*, *psbI*, *trnS–GCU* and *trnG–UCC*), resulting in a c.11-kb duplicated fragment.

Two exceptions to the IR/LSC boundary shift were *Begonia ulmifolia* Willd., which had a secondary IR expansion from IRb to LSC including a partial *rps19* sequence (153 bp) (see Figure 2), and *B. microsperma* Warb., which had an addition 67-bp IR expansion from IRa to SSC. Additionally, we found a 213-bp inversion in the *ndhF–rpl32* spacer of *Begonia fenicis* (Appendix figure).

Simple sequence repeats and tandem repeat analyses

In the present study, the number of SSRs found within *Begonia* plastomes ranged from 119 (*B. henryi*) to 171 (*B. ulmifolia*) (mean = 149.7), whereas 115 SSRs were found in *Hillebrandia sandwicensis* (Figure 3), compared with 99 SSRs in *Arabidopsis*, 92 in *Ampelosicyos*

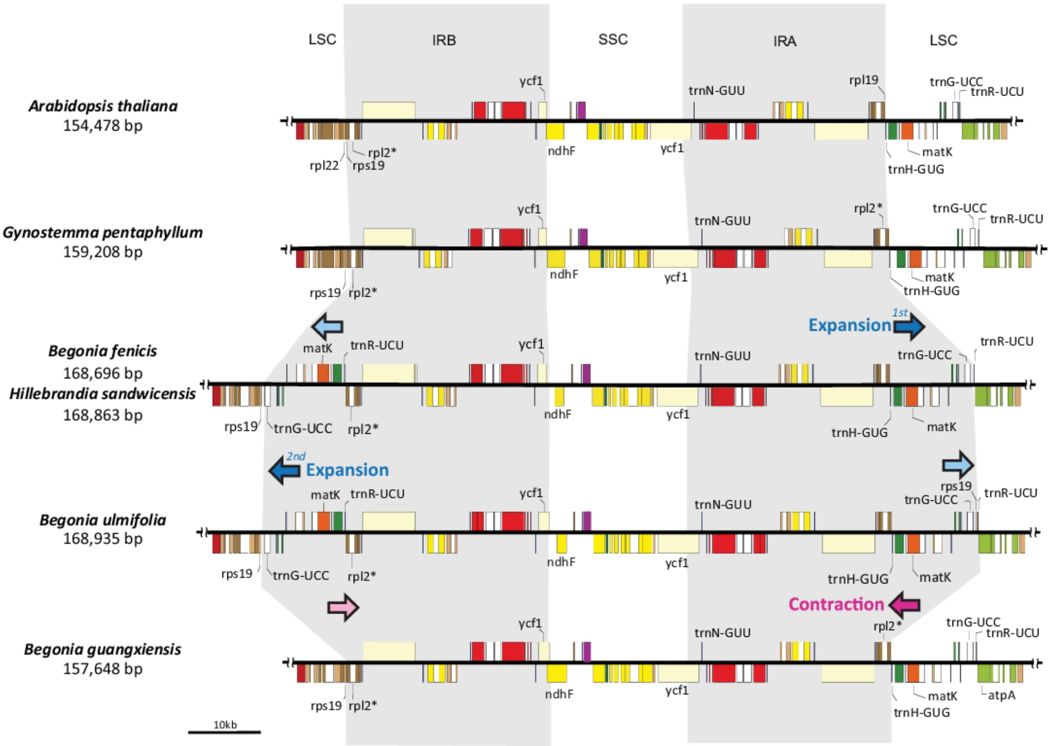


Figure 2. Comparison of the boundaries of large single copy (LSC), small single copy (SSC) and inverted repeat (IR) regions in *Arabidopsis thaliana*, *Gynostemma pentaphyllum*, three *Begonia* species and *Hillebrandia sandwicensis*. In contrast to the plastome sequences of *Arabidopsis thaliana* and *Gynostemma pentaphyllum*, those of most *Begonia* sampled in the present study (here represented by *B. fenicis*) and *Hillebrandia sandwicensis* have identical plastome structures, with an IR expansion (labelled '1st Expansion') resulting in an approximately 11-kb duplicated fragment. *Begonia ulmifolia* has another 153-bp IR expansion (labeled '2nd Expansion') from IRb to LSC and including a partial *rps19* sequence. The published plastome of *Begonia guangxiensis* (Dong *et al*., 2019) includes contracted IRs, similar to those of *Arabidopsis thaliana* and *Gynostemma pentaphyllum*. The grey box represents the range of IR. An asterisk indicates that the gene carries an intron.

humblotii and 92 in *Gynostemma pentaphyllum*. Among all SSRs, the most abundant type were mononucleotide repeats, which accounted for 77.0% (mean n = 115.3) of the total SSRs, followed by dinucleotide (13.4%, 20.1), tetranucleotide (6.7%, 10.0), trinucleotide (2.3%, 3.4), pentanucleotide (0.3%, 0.5) and hexanucleotide (0.2%, 0.1) repeats (Table 3A). Most SSRs were in the LSC region (60.4%; mean of number = 90.3), 20.8% (31.3) in the SSC regions and 18.8% (28.1) in the IR (see Figure 3B).

Among these SSRs, 41 different SSR types were found. One mononucleotide repeat unit (A/T), two dinucleotide repeat units (AT/AT, AG/CT) and one tetranucleotide repeat (AAAT/

Table 3. Potential DNA barcode sequences

Marker	Length range (bp)	GC content (%)	Length of variable site (bp)	Proportion of variable site (%)	Parsimony-informative site (bp)	Proportion of parsimony-informative site (%)
trnE–trnT spacer	536–1230	21.4	485	33.7	260	18.1
rbcL–accD spacer	628–715	27.9	261	32.2	133	16.4
ycf1–ndhF spacer	970–1045	25.4	311	29.0	171	16.0
ndhF–rpl32 spacer	695–1016	20.6	473	39.3	277	23.0
rps15–ycf1 spacer	371–452	19.9	179	32.7	102	18.6
ycf1-partial	4249–4544	26.0	1630	34.7	956	20.4

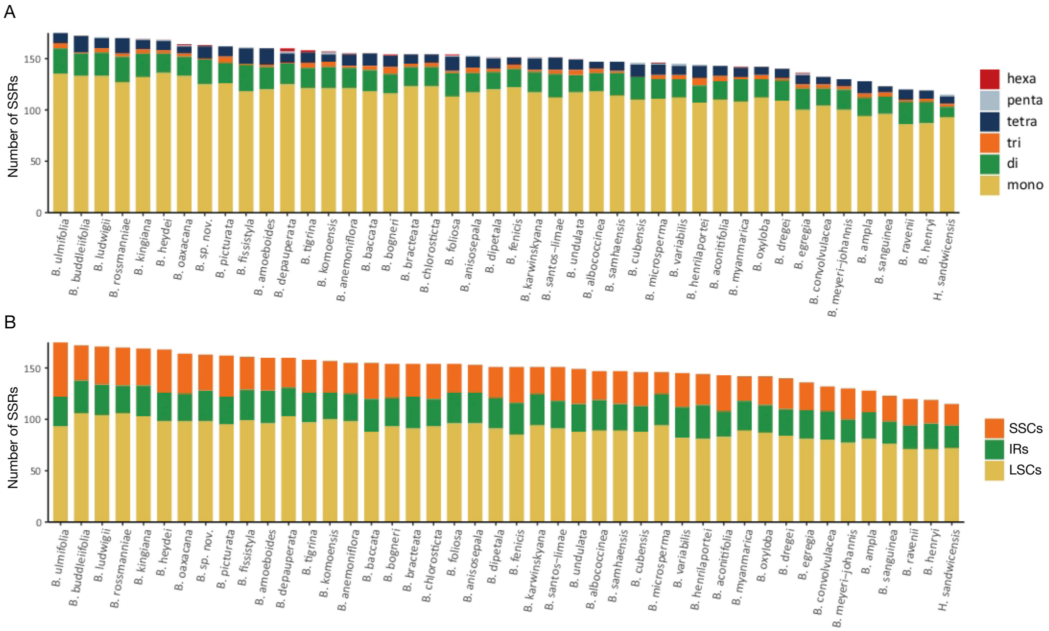


Figure 3. Comparison of the simple sequence repeats (SSRs) in 44 plastomes of Begoniaceae: A, number of SSR types detected in each plastome; B, distribution of SSRs across small single copies (SSCs), inverted repeats (IRs) and large single copies (LSCs).

ATTT) were found in all 44 samples (Figure 4). The A/T repeat was the most abundant (76.5%, *n* = 114.6), followed by the AT/AT repeat (11.8%, 17.7). Of the dinucleotide repeats, *Hillebrandia sandwicensis* had the fewest (*n* = 10, 17–26 in *Begonia*), and the AC/GT repeat unit was detected in all samples except that of *B. chlorosticta* Sands, *B. meyeri-johannis*, *B. oxyloba* Welw. ex Hook.f. and *Hillebrandia sandwicensis*.

In trinucleotide repeats, the AAT/ATT repeat unit was found in all samples except that of *Begonia cubensis* Hassk. The other trinucleotide repeats were found only in African species

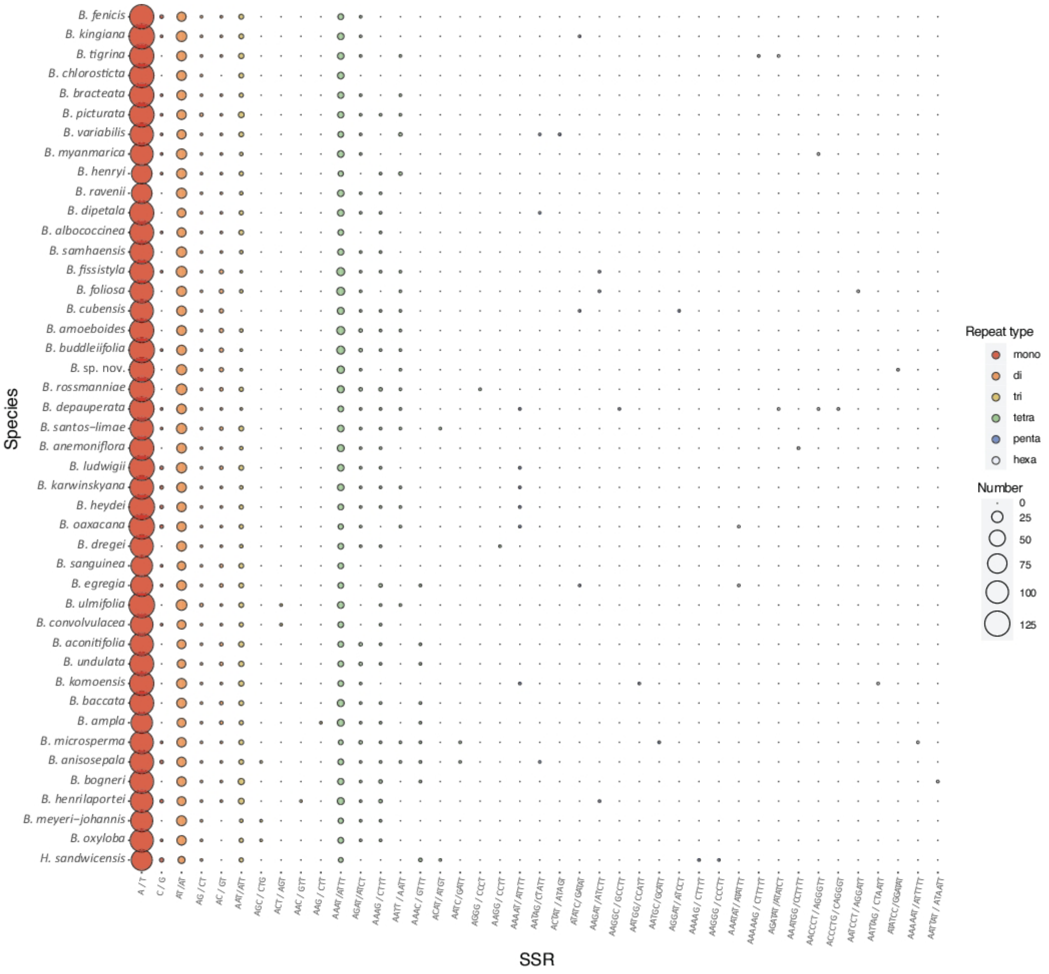


Figure 4. Distribution of the 41 types of simple sequence repeat (SSR) units across 44 plastomes of Begoniaceae. The colours indicate the SSR types, and the size of each circle corresponds with the number of SSRs. The species are ordered according to their phylogenetic relatedness, as shown in Figure 7.

(*Begonia ampla* Hook.f., *B. anisosepala*, *B. henrilaportei* Scherber. & Duruiss., *B. meyeri-johannis* and *B. oxyloba*) and in species from Neotropical clade 2 (*B. convolvulacea* and *B. ulmifolia*). Most pentanucleotide and hexanucleotide repeats were usually species-specific; for example, two repeat units, AAAAG/CTTTT and AAGGG/CCCTT, were detected only in *Hillebrandia sandwicensis*.

The total number of long repeats in Begoniaceae plastomes was between 75 (*Begonia ludwigii* Irmsch.) and 26 (*B. sanguinea* Raddi) (Figure 5). The length of long repeats ranged from 30 to 178 bp, with 30–39 bp being the most common (59.6%–88.6%), of which the

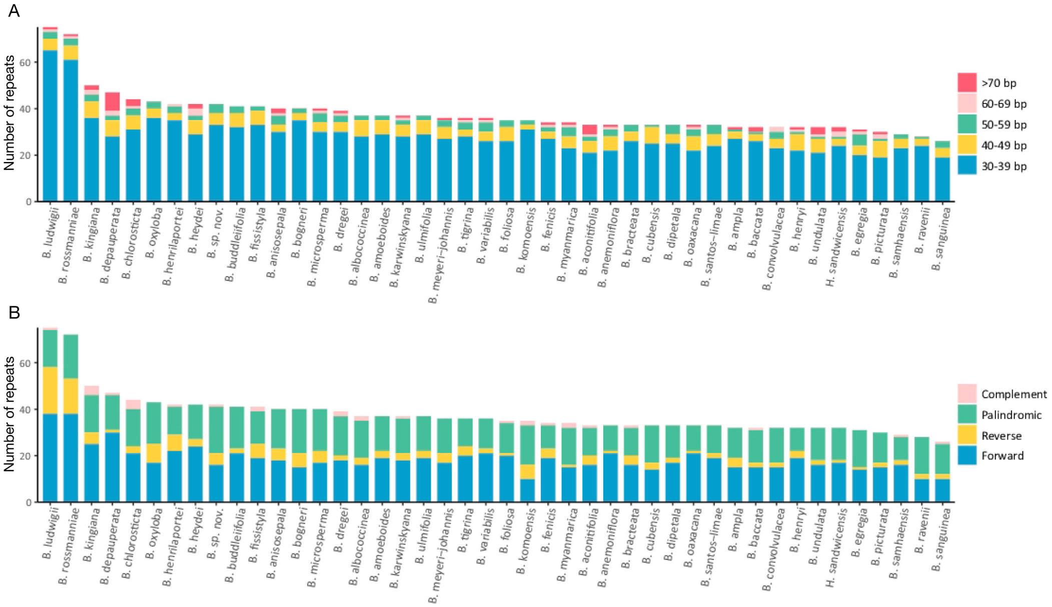


Figure 5. Analysis of long repeats in 44 plastomes of Begoniaceae: A, distribution and lengths of long repeats; B, numbers of four types of repeat. IR, inverted repeat; LSC, large single copy; SSC, small single copy.

most frequent were 30 bp, 31 bp, 35 bp, 34 bp and 32 bp long. The number of forward repeats varied between 10 (*Begonia komoensis* Irmsch., *B. ravenii* C.I Peng & Y.K.Chen and *B. sanguinea*) and 38 (*B. ludwigii*); the number of reverse repeats varied from one (*B. egregia* N.E.Br., *B. myanmarica* C.I Peng & Y.D.Kim and *Hillebrandia sandwicensis*) to 20 (*B. ludwigii*), palindromic repeats from 10 (*B. fenicis*) to 20 (*Begonia* sp. nov. sect. *Ruizopavonia*), and complement repeats from zero (*B. anisosepala*, *B. bogneri* Ziesenh., *B. convolvulacea*, *B. cubensis*, *B. egregia*, *B. microsperma*, *B. picturata* Yan Liu, S.M.Ku & C.I Peng, *B. ravenii*, *B. undulata* Schott and *H. sandwicensis*) to four (*B. chlorosticta*). The composition of long repeat sequences also varies among species, with forward repeats more common in 32 species, and palindromic repeats more frequent in 11 species and equally common in *Begonia convolvulacea*.

Sequence divergence and nucleotide diversity

The mean nucleotide diversity (π) of plastomes was estimated to be 0.0142 in *Begonia*, with a range of 0.0002–0.083. The SSC region had the highest mean nucleotide diversity (π = 0.0345), followed by the LSC region (π = 0.0202) and the IR region (π = 0.0041). Based on the sliding window analysis (Figure 6), six regions, including five spacers (*trnE–trnT*, *rbcL–trnD*, *ycf1–ndhF*, *ndhF–rpl32* and *rpl15–ycf1*) and one gene (the first 4500 bp of *ycf1*),

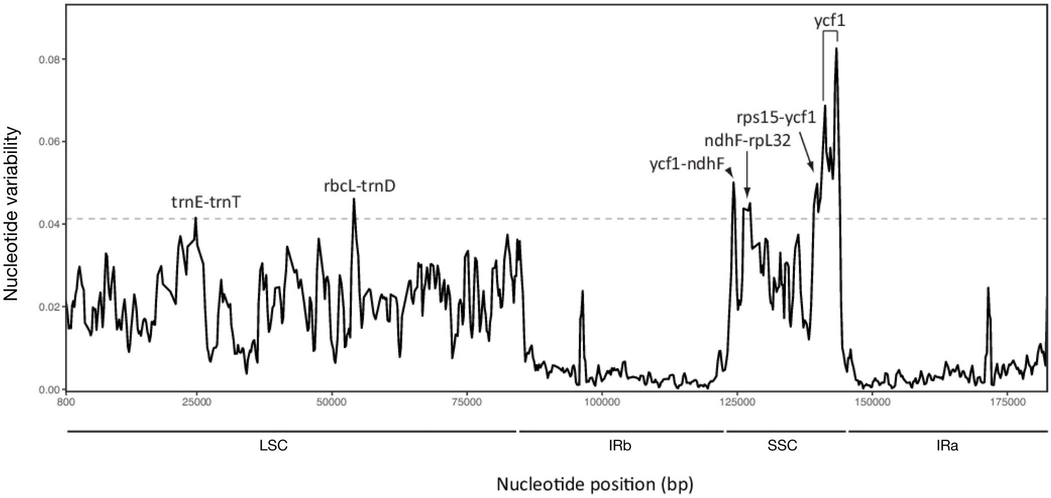


Figure 6. Sliding-window analysis of the complete plastomes of 43 *Begonia* species. The dashed line represents the π value that is higher than 95% of all values (π = 0.0413) in our data.

were identified as the most variable regions. Among the six regions, the *ndhF–rpl32* spacer contained the highest proportion of variable and parsimony-informative sites (39.3% and 23.0%), followed by the first 4500 bp of the *ycf1* gene (34.7% and 20.4%) and the *rps15–ycf1* spacer (32.7% and 18.6%) (see Table 3).

Phylogenetic analyses

The total length of the plastome alignment for the phylogeny (including SSC, LSC, and one IR) was 150,446 bp, of which 32,949 bp were variable sites (21.9%) and 18,058 bp were parsimony-informative sites (12.0%). The phylogenetic analyses constructed by the ML and BI methods showed identical topologies (Figure 7). Begoniaceae was supported as a monophyletic group (BS = 100, PP = 1), and *Begonia* formed a well-supported clade (BS = 100, PP = 1) sister to the monotypic *Hillebrandia*. The major clades recovered in the plastome tree were mostly congruent with the geographical distribution (see Figure 7).

All African species except for *Begonia dregei* Otto & A.Dietr. and *B. samhaensis* M.Hughes & A.G.Mill. formed four highly supported clades and occupied the most basal lineage of *Begonia* (BS = 100, PP = 1); the clades were *Begonia* sect. *Exalabegonia*, Malagasy *Begonia* (MB), yellow-flowered African *Begonia* (YFAB) and fleshy-fruited African *Begonia* (FFAB), congruent with the groupings of Moonlight *et al.* (2018). A similar pattern was also found for Asian *Begonia*, which consisted of three major clades corresponding to Asian clade D (BS = 100, PP = 1), Asian clade C and EDAB (early-diverging Asian *Begonia*) (BS = 100, PP = 1 in both groups), based on the nomenclature of Moonlight *et al.* (2018). Species from

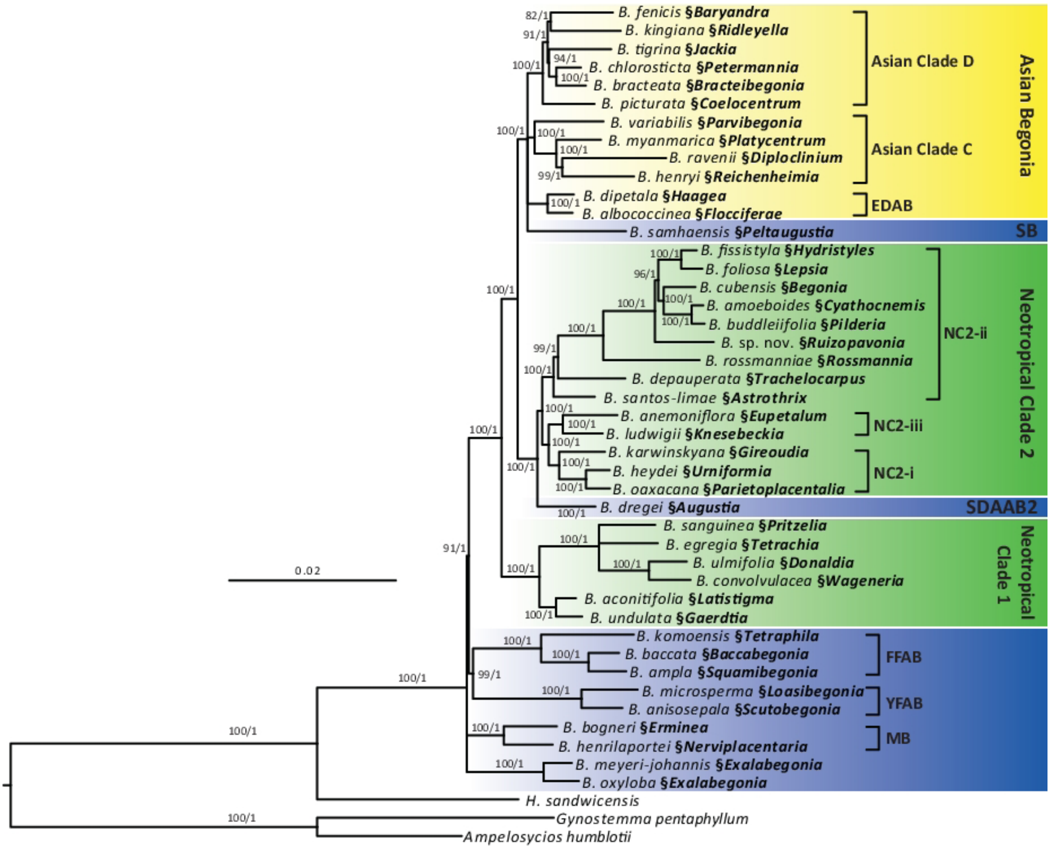


Figure 7. Phylogenomic tree of Begoniaceae generated from RAxML maximum likelihood (ML) analysis based on plastome sequences (small single copies, large single copies and single inverted repeat). Bootstrap values ≥ 80 and posterior probability values = 1 from ML and Bayesian inference analyses, respectively, are shown on the branches. Colours indicate the geographical distribution of the species: blue, Africa; green, the Americas; yellow, Asia. Naming of the clades is based on the system of Moonlight *et al.* (2018). Abbreviated clade names: EDAB, early-diverging Asian *Begonia*; FFAB, fleshy-fruited African *Begonia*; MB, Malagasy *Begonia*; NC2, Neotropical clade 2; SB, Socotran *Begonia*; SDAAB, seasonally dry–adapted African *Begonia*; YFAB, yellow-flowered African *Begonia*. §, Section.

the Americas formed two major clades: Neotropical clade 1 (NC1) (BS = 100, PP = 1) and Neotropical clade 2 (NC2) (BS = 100, PP = 1). In NC2, there were three major clades, namely NC2-i, NC2-ii and NC2-iii. *Begonia dregei* from Africa was the most basal lineage of NC2. All sampled Asian *Begonia* and *B. samhaensis* (SB) formed a clade (BS = 100, PP = 1) sister to NC2.

Discussion

Plastome structure of Begoniaceae

The length of Begoniaceae plastomes (mean = 169,426 bp) is greater than those of most land plants, which typically range from 120 to 160 kb (Twyford & Ness, 2017). Compared with most angiosperms with a typical IR length of c.25 kb (Ruhlman & Jansen, 2014), the length of the IR region of Begoniaceae is increased to 37.5 kb, suggesting a large-scale of IR expansion in this family. IR expansion and contraction often leads to variation in plastome size among different plant groups and has been suggested as the cause of gene order changes (Jansen & Ruhlman, 2012). Similarly, the increase in total plastome size observed in Begoniaceae could be explained by the expansion of IRs (Asaf *et al.*, 2016; Xu *et al.*, 2017; Li & Zheng, 2018).

We found the IR/SSC boundary to be highly conserved in Begoniaceae; it is located between *trnN–GUU* and *ycf1* on IRa/SSC and within the 5' end of *ycf1* on SSC/IRb. The last full-length gene in the IR at the IRa/SSC boundary is *trnN–GUU*, a pattern similar to that observed in the most other land plants (Raubeson *et al.*, 2007; Zhu *et al.*, 2016). Additionally, except for *Begonia guangxiensis* C.Y.Wu and *B. ulmifolia*, the IRb/LSC boundary is also highly conserved in Begoniaceae, with the IR expansion of a duplicated fragment of 10 genes (i.e. *trnH–GUG*, *psbA*, *matK*, *trnK–UUU*, *rps16*, *trnQ–UUG*, *psbK*, *psbI*, *trnS–GCU* and *trnG–UCC*).

Generally, the IR/LSC boundaries of non-monocot angiosperms occur between *rpl2* and *rps19* (type I) or in *rps19* (type II) on the IRb/LSC side, and between *rpl2* and *trnH–GUG* on the IRa/LSC side (Wang *et al.*, 2008). However, many studies have shown that IRs of angiosperm plastomes fluctuate greatly in size because of the expansion of the IRb/LSC boundary (Downie & Jansen, 2015) from *rps19* to *rbcL*. For example, IRbs expanded to *rpl22* in *Halenia elliptica* D.Don (Gentianaceae; Zhang *et al.*, 2020), *rps3* in *Citrus limon* (L.) Osbeck (Rutaceae; Khan *et al.*, 2019), *rpl16* in *Pachysandra* Michx. (Buxaceae; Sun *et al.*, 2016), *pet*D in several species of *Amphilophium* Kunth (Bignoniaceae; Thode & Lohmann, 2019), *pet*B in *Anemopaegma prostratum* DC. (Bignoniaceae; Thode & Lohmann, 2019), between *psbB* and *clpP* in *Mahonia* Nutt. and *Berberis* L. (Berberidaceae; Kim & Jansen, 1994), *clpP* in *Nicotiana acuminata* (Graham) Hook. (Solanaceae; Shen *et al.*, 1982; Goulding *et al.*, 1996), and *rbcL* in *Pelargonium* × *hortorum* L.H.Bailey (Geraniaceae; Chumley *et al.*, 2006). In comparison, there are relatively few examples of IRa/LSC expansion in non-monocot angiosperms. In most cases, IRa either expands to *trnH–GUG* or contains a *trnH–rps19* gene cluster that is similar to the IRa/LSC boundary in most monocots (Wang *et al.*, 2008). *Strobilanthes cusia* (Nees) Kuntze (Acanthaceae) is another case of IRa-to-LSC expansion, resulting in the inclusion of two copies of the *trnH–GUG* gene and two partial *psbA* genes in its plastome (Chen *et al.*, 2018).

To the best of our knowledge, the IRa-to-LSC expansion in Begoniaceae is the longest among the land plants. The unique IR expansion in *Begonia* could explain the failure of the attempt by Harrison *et al.* (2016) to amplify *rpl2–rps19–rpl22* genes in 16 *Begonia* species by means of PCR using custom primers. Additionally, this expansion appears to have occurred after the divergence of Begoniaceae and Cucurbitaceae and before the split between *Begonia* and *Hillebrandia*. Gene duplication caused by IR is suggested to be an important driving force in the evolution of plastomes, leading to the increases of genes and gene complexity, which are two significant factors correlated to origin of genomic and organismal complexity (Xiong *et al.*, 2009).

Among the 10 duplicated genes contributing to the IR expansion, *matK* has been suggested to be important for the splicing of RNAs with essential roles for translational apparatus and plant cell survival (Zoschke *et al.*, 2010). In most land plants, *matK* is a single copy and one of the fastest evolving genes in protein-encoding regions of the plastome (Wolfe, 1991), therefore it has been commonly used in systematic and evolutionary studies as a core species barcode (Hollingsworth *et al.*, 2009). However, *matK* has a rate deceleration in *Begonia*, causing it to be less useful in phylogenetic studies (Daniel Thomas & Mark Hughes, unpublished data). This finding could be explained by the idea that the sequences located in the IR usually have lower substitution rates, because the two identical copies of IR provide a template for error correction when a mutation occurs in one of the copies (Weng *et al.*, 2017). Additionally, a recent study demonstrated that the ectopic insertion of *matK* could lead to variegated cotyledons in tobacco (Qu *et al.*, 2018). *Begonia* species are well known for their various natural foliar variegation patterns with uneven distribution of pigmentation and silvery spots (Sheue *et al.*, 2012). Future studies on the correlation between *matK* function and variegation in *Begonia* may allow us to understand the mechanism and evolutionary process underlying the unique IRa-to-LSC expansion and leaf variegation diversity in this megadiverse genus.

Both *Hillebrandia* and the majority of *Begonia* taxa are characterised by a unique IRa-to-LSC expansion in their plastomes, therefore the shifts of the IR boundary observed in *B. ulmifolia* and *B. guangxiensis* are more likely to have occurred independently, because these two species are distantly related (Moonlight *et al.*, 2018). We further identified a secondary expansion (153 bp), from IRb to LSC including a partial *rps19* sequence, in the plastome of *Begonia ulmifolia*. Rather than possessing the IRa-to-LSC expansion typical in the majority of Begoniaceae, the published plastome of *Begonia guangxiensis* (Dong *et al.*, 2019) has contracted IRs, similar to those of *Gynostemma pentaphyllum* (Cucurbitaceae) and most angiosperms, and which we did not observe in any other Begoniaceae plastomes.

Begonia guangxiensis, distributed in the limestone karsts of the Sino-Vietnamese region, is classified in *Begonia* sect. *Coelocentrum* (Wu & Ku, 1997; Chung *et al.*, 2014). However, the results of our study show that *Begonia picturata*, also in *Begonia* sect. *Coelocentrum*, has a similar plastome structure to that of most Begoniaceae with an IRa-to-LSC expansion. Moreover, this typical IR expansion of Begoniaceae has also been observed in other species of *Begonia* sect. *Coelocentrum* (Y.-H. Tseng *et al.*, unpublished data). Therefore, the IR contraction of *Begonia guangxiensis* and the secondary IR expansion of *B. ulmifolia* represent the species-specific cases during Begoniaceae evolution, confirming that the IR/LSC boundary is not static but could be affected by a dynamic and random process that allows expansion and contraction of the IR (Goulding *et al.*, 1996).

In the present study, we detected only one inversion in our Begoniaceae plastomes, an approximately 210-bp inversion in the *ndhF–rpl32* spacer of *Begonia fenicis* (sect. *Baryandra*). By scrutinising the alignment of the *ndhF–rpl32* sequence across different *Begonia* species, we found that the presence of the inversion could be unique to *Begonia* sect. *Baryandra*. More unpublished plastome data for *Begonia* sect. *Baryandra* (L. W. Tsai *et al.*, unpublished) support this hypothesis. Rearrangements such as inversions in the plastomes are considered useful markers with which to infer evolutionary relationships of land plants (Doyle *et al.*, 1992). Therefore, the unique inversion detected in *Begonia* sect. *Baryandra* might be a powerful marker for use in identifying species from this section, a species-rich lineage from around the Philippine Archipelago (Hughes *et al.*, 2015, 2018). We note that the *ndhF–rpl32* spacer has been widely used in previous phylogenetic analyses of *Begonia*, especially at the species and sectional levels (Thomas *et al.*, 2012; Hughes *et al.*, 2015, 2018). It is necessary to be cautious in aligning this inversion of the *ndhF–rpl32* spacer in *Begonia* sect. *Baryandra* when doing phylogenetic analysis and estimating the nucleotide diversity.

Simple sequence repeats and dispersed repeats in Begonia

Simple sequence repeats are frequently observed in plastomes, which are of particular interest in studies of evolution, population genetics and genome polymorphism (Ebert & Peakall, 2009; Qi *et al.*, 2016). Several nuclear SSRs in *Begonia* have been identified in previous studies (Hughes *et al.*, 2003; Twyford *et al.*, 2013a; Chan *et al.*, 2014; 2015; Tseng *et al.*, 2017); however, the results of only one study based on plastid-derived microsatellite markers have been published so far (Twyford *et al.*, 2013b). In our study, we successfully detected approximately 150 (range, 115–171) plastome-derived SSRs per species, which could be useful for further population genetic and biogeographical analyses of Begoniaceae.

Compared with nuclear microsatellites, plastid SSRs are generally more suitable for studies of seed dispersal patterns, species distribution changes, genetic drift, and assessment of haploid distribution across similar geographical areas (Twyford *et al.*, 2013b). Caution is nevertheless necessary, because we find that some SSRs do not appear consistently across the Begoniaceae and therefore may not be useful in reflecting phylogenetic relatedness. For example, the repeat of ACAT/ATGT was found only in *Hillebrandia sandwicensis* and *Begonia santos-limae* Brade, and the repeat of AGAAT/ATTCT was detected only in *B. depauperata* Schott from the Americas and *B. tigrina* Kiew from Asia. The independent occurrence of SSRs in Begoniaceae implies that the dynamics of plastomes may affect the successful rate of interspecific transferability of SSRs for further population genetics analyses.

Several studies have demonstrated a positive correlation between dispersed repeats and rearrangements (Lee *et al.*, 2007; Guisinger *et al.*, 2011; Weng *et al.*, 2014), leading to the suggestion that long repeats (dispersed repeats) are a major factor promoting plastome rearrangement in land plants. However, contrasting cases have been reported for *Coffeea* DC. (Rubiaceae; Samson *et al.*, 2007) and *Daucus* L. (Apiaceae; Ruhlman *et al.*, 2006), in which there is no correlation between the number and type of repeats and the propensity for genome rearrangements. In the present study, we revealed high variation in the number, type and composition of the long repeat sequences in the Begoniaceae; however, plastome gene order and content were found to be highly conserved, with no rearrangement having been detected in most species of the family. More comprehensive genomic studies are needed to explore these repeat elements in Begoniaceae.

DNA tandem repeats are another popular molecular marker in addition to important genomic elements from the evolutionary and functional perspectives (Jernigan & Bordenstein, 2015; Gymrek *et al.*, 2016; Zhao *et al.*, 2018). Nearly all detected mutations in the spontaneous plastome mutants could be associated with repetitive elements (Massouh *et al.*, 2016), suggesting that tandem repeats play important roles in plastid genome variation between closely related species (Li *et al.*, 2019). Recent work by Picart-Picolo *et al.* (2020) provides evidence of satellite DNA changes as modifiers of genome structure and stability that can trigger gene duplication and structural variations carrying changes in expression patterns. Moreover, nuclear satellite DNA sequences are rapidly evolving sequences that may cause reproductive barriers between organisms and promote speciation (Garrido-Ramos, 2015).

Our analysis shows that *Hillbrandia sandwicensis* has fewer SSRs (*n* = 115) and repeats (*n* = 32) in its plastome than most studied *Begonia* species (mean *n* = 149.7 and 37.9, respectively), suggesting that *Begonia* species have more repetitive and dynamic plastomes than those of *H. sandwicensis*. Although nuclear–plastid DNA exchange is a common event in plants (Gao *et al.*, 2014) and has been recently described in other Cucurbitales (Cui *et al.*, 2021), further analysis on nuclear genome repeats would be required to elucidate whether the amount of satellite DNA in plastome genomes can be linked with more dynamic and repetitive nuclear genomes, and the role genome dynamics might play in the evolution of *Begonia*.

Potential DNA barcodes for Begonia

Comparative genomic analyses of complete plastome sequences are necessary for developing variable DNA barcode regions as potential molecular markers for species identification (Nock *et al.*, 2011). In our study, we identified six highly variable regions, namely, *trnE–trnT*, *rbcL–trnD*, *ycf1–ndhF*, *ndhF–rpl32*, *rpl15–ycf1* and part of the *ycf1* gene, as potential DNA barcodes in *Begonia*. However, the results of an analysis of 24 plastid regions from 16 *Begonia* species (Harrison *et al.* (2016) suggest *rpoB–psbD* (a sequence between *rpoB* and *psbM*) and a partial sequence of *ndhI–ndhG* (804 bp) as candidate markers for phylogenetically informative plastid regions for use in low-level studies in *Begonia*. Compared with the nucleotide diversity (π) of these two sequences in our study, π values for the *rpoB–psbM* sequence (0.0264) and *ndhI–ndhG* sequence (0.0305) are lower than that of our six candidate markers (π > 0.04). The difference might be due to the different taxonomic sampling in the two studies. Harrison *et al.* (2016) sampled 16 species in nine sections, eight of which were from *Begonia* sect. *Gireoudia*, whereas in our study we sampled one species per section but included more sections (43 species in 42 sections). The *ndhF–rpl32* spacer, with the highest proportion of variable and parsimony-informative sites among our six candidate markers, has been shown to resolve successfully the phylogenetic framework of *Begonia* at both sectional and species level (Thomas *et al.*, 2012; Hughes *et al.*, 2015, 2018).

After investigating the reliability and effectiveness of five other candidate markers for DNA barcodes for use in phylogenetic studies in *Begonia*, we expect that the candidate markers reported here could be useful references for further studies on genetic diversity assessment, phylogenetics and population genetics in *Begonia* with more comprehensive taxon sampling.

Plastome phylogenomic relationship of Begonia

The pantropical genus *Begonia* is the fifth largest genus of flowering plants, comprising more than 2000 described species. Establishing a robust phylogeny as a backbone of this megadiverse genus will provide a fundamental framework for further taxonomic, ecological and evolutionary studies. Based on three plastid gene sequences (Moonlight *et al.*, 2018) and whole-plastome sequences (Shui *et al.*, 2019), two drastically different infrageneric classification systems of *Begonia* have been proposed. These two conflicting systems are mainly due to different views on monophyly versus paraphyly, taxon versus gene sampling, the importance of morphology, and the understanding of phylogenetic conflicts (Shui *et al.*, 2019). Although in our study we analysed only 42 of the 70 known sections in *Begonia*, and each section was represented by only one species, the robust plastome phylogeny presented here is still useful in confirming the major groups and highlighting potential classification questions in *Begonia*.

As in two previous studies (Moonlight *et al.*, 2018; Shui *et al.*, 2019), our results show that *Begonia* can be divided into four major groups related to geographical distribution: an early-diverging group of African species, two clades from the Americas, and one Asian clade. Another congruence is the placement of African species, *Begonia dregei*, as sister to species of the clade NC2 (Moonlight *et al.*, 2018; Shui *et al.*, 2019). However, the early-diverging African group is not congruent between these three phylogenies. In our study, the MB (Malagasy *Begonia*) clade and the *Begonia* sect. *Exalabegonia* clade are unresolved as sister to the rest of *Begonia*, but in the studies by Moonlight *et al.* (2018) and Shui *et al.* (2019), the YFAB (yellow-flowered African *Begonia*) and a larger African clade (Group 1) is the sister lineage, respectively.

Another conflict regards the phylogenetic relationships between the three clades in Neotropical clade 2. Our results show that the NC2-i clade is sister to the NC2-iii clade rather than the NC2-ii clade, as in Moonlight *et al.* (2018). Additionally, conflicting phylogenetic replacements are also apparent within the NC2-ii clade (Moonlight *et al.*, 2018). In Asian clade D, the relationships between *Begonia* sects *Baryandra*, *Jackia*, *Ridleyella*, *Petermannia* and *Bracteibegonia* differ slightly from those described by Moonlight *et al.* (2018), in which the relationships, although resolved, were not well supported. A further sampling of five species of *Begonia* sect. *Petermannia*, by Shui *et al.* (2019), showed that the section forms a grade with *Begonia* sects *Ridleyella* and *Baryandra* nested within. Based on the concept of monophyly-based taxonomy, Shui *et al.* (2019) proposed a much expanded *Begonia* sect. *Petermannia*, including all the five sections in Asian clade D. Earlier work with better taxon sampling has shown that there is evidence for two distinct clades of species in *Begonia* sect. *Petermannia* that may provide clues for subdividing the section into more manageable units (Thomas *et al.*, 2011; Moonlight *et al.*, 2018).

The supported differences between the phylogenies in the present study, and those by Moonlight *et al.* (2018) and Shui *et al.* (2019), are perhaps surprising given that they are all based on the plastid genome. The differences could be due to alternative alignments of hypervariable spacer regions, or long-branch attraction when taxon sampling is poor (Stefanović *et al.*, 2004; Bergsten, 2005). Although our plastome sequence data have allowed us to recover a well-resolved phylogeny in *Begonia*, future work must combine increased genomic sampling with a much denser taxon sampling to further explore areas of uncertain topology and to circumscribe problematic sections, such as *Petermannia*.

Conclusions

Our comparative analyses of 44 Begoniaceae plastomes provide important insights into the structure and evolution of the Begoniaceae plastome. The highly conserved plastomes of the Begoniaceae have a unique IR expansion from IRa to LSC, with a duplicated fragment from the *trnH–GUG* gene to the *trnR–UCU* gene, probably the first known case in land plants. Based on the analyses of SSRs and repeats, our results suggest that *Begonia*, as a species-rich genus, has a more repetitive and dynamic plastome than that of its sister and monotypic genus *Hillebrandia*. Moreover, the robust plastome phylogeny in this study provides a framework for further taxonomic, evolutionary and biogeographical research. Nevertheless, additional and comprehensive sampling is required to further investigate the evolution of the plastome and address the infrageneric classification of *Begonia*, especially with respect to conflicting taxonomic treatments.

Acknowledgements

The authors thank the staff of the Experimental Greenhouse and NGS High Throughput Genomic Core of BRCAS. Thanks also to Hsun-An Yang and Tsui-Ya Liu for confirming the information on species distributions.

This study was supported by funding from Academia Sinica, Taiwan, to K.-F.C. We thank the Peruvian authorities for granting collecting permits (#08-2014-SERFOR).

ORCID iDs

Y.-H. Tseng https://orcid.org/0000-0002-8166-5690
C. L. Hsieh https://orcid.org/0000-0002-3342-3654
L. Campos-Domínguez https://orcid.org/0000-0002-8998-3394
A. Q. Hu https://orcid.org/0000-0001-9564-878X
C. C. Chang https://orcid.org/0000-0002-3035-7348
Y. T. Hsu https://orcid.org/0000-0002-6862-4747
C. A. Kidner https://orcid.org/0000-0001-6426-3000
M. Hughes https://orcid.org/0000-0002-2168-0514
P. W. Moonlight https://orcid.org/0000-0003-4342-2089
Y. C. Wang https://orcid.org/0000-0002-6544-3473
Y. T. Wang https://orcid.org/0000-0003-3668-219X
S. H. Liu https://orcid.org/0000-0002-5429-6869
D. Girmansyah https://orcid.org/0000-0002-3096-8763
K.-F. Chung https://orcid.org/0000-0003-3628-2567

References

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ [Accessed 1 May 2021.]

Asaf S, Khan AL, Khan AR, Waqas M, Kang SM, Khan MA, Lee SM, Lee IJ. 2016. Complete chloroplast genome of *Nicotiana otophora* and its comparison with related species. Frontiers in Plant Science. 7:e843. https://doi.org/10.3389/fpls.2016.00843

Asaf S, Khan AL, Lubna, Khan A, Khan A, Khan G, Lee IJ, Al-Harrasi A. 2020. Expanded inverted repeat region with large scale inversion in the first complete plastid genome sequence of *Plantago ovata*. Scientific Reports. 10(1):3881. https://doi.org/10.1038/s41598-020-60803-y

Beier S, Thiel T, Munch T, Scholz U, Mascher M. 2017. MISA-web: a web server for microsatellite prediction. Bioinformatics. 33(16):2583–2585. https://doi.org/10.1093/bioinformatics/btx198

Bergsten J. 2005. A review of long-branch attraction. Cladistics. 21(2):163–193. https://doi.org/10.1111/j.1096-0031.2005.00059.x

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 30(15):2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Borowiec ML. 2016. AMAS: a fast tool for alignment manipulation and computing of summary statistics. Peer J. 4:1660. https://doi.org/10.7717/peerj.1660

Chan PP, Lowe TM. 2019. tRNAscan-SE: searching for tRNA genes in genomic sequences. Methods in Molecular Biology. 1962:1–14. https://doi.org/10.1007/978-1-4939-9173-0_1

Chan YM, Lee CT, Tnah LH, Lee SL. 2014. Novel microsatellite markers for *Begonia maxwelliana* and transferability to 23 *Begonia* species of Peninsular Malaysia. Biochemical Systematics and Ecology. 57:159–163. https://doi.org/10.1016/j.bse.2014.08.004

Chan YM, Twyford AD, Tnah LH, Lee CT. 2015. Characterisation of EST-SSR markers for *Begonia maxwelliana* (Begoniaceae) and cross-amplification in 23 species from 7 Asian sections. Scientia Horticulturae. 190:70–74. https://doi.org/10.1016/j.scienta.2015.04.012

Chen HM, Shao JJ, Zhang H, Jiang M, Huang LF, Zhang Z, Yang D, He M, Ronaghi M, Luo X, Sun B, Wu WW, Liu C. 2018. Sequencing and analysis of *Strobilanthes cusia* (Nees) Kuntze chloroplast genome revealed the rare simultaneous contraction and expansion of the inverted repeat region in angiosperm. Frontiers in Plant Science. 9:324. https://doi.org/10.3389/fpls.2018.00324

Cheng JW, Zhao ZC, Li B, Qin C, Wu ZM, Trejo-Saavedra DL, Luo XR, Cui JJ, Rivera-Bustamante RF, Li SC, Hu KL. 2016. A comprehensive characterization of simple sequence repeats in pepper genomes provides valuable resources for marker development in *Capsicum*. Scientific Reports. 6:18919. https://doi.org/10.1038/srep18919

Chumley TW, Palmer JD, Mower JP, Fourcade HM, Calie PJ, Boore JL, Jansen RK. 2006. The complete chloroplast genome sequence of *Pelargonium* × *hortorum*: organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. Molecular Biology and Evolution. 23(11):2175–2190. https://doi.org/10.1093/molbev/msl089

Chung KF, Leong WC, Rubite RR, Repin R, Kiew R, Liu Y, Peng CI. 2014. Phylogenetic analyses of *Begonia* sect. *Coelocentrum* and allied limestone species of China shed light on the evolution of Sino-Vietnamese karst flora. Botanical Studies. 55:1. https://doi.org/10.1186/1999-3110-55-1

Clement WL, Tebbitt MC, Forrest LL, Blair JE, Brouillet L, Eriksson T, Swensen SM. 2004. Phylogenetic position and biogeography of *Hillebrandia sandwicensis* (Begoniaceae): a rare Hawaiian relict. American Journal of Botany. 91(6):905–917. https://doi.org/10.3732/ajb.91.6.905

Cui H, Ding Z, Zhu Q, Wu Y, Qiu B, Gao P. 2021. Comparative analysis of nuclear, chloroplast, and mitochondrial genomes of watermelon and melon provides evidence of gene transfer. Scientific reports. 11:1595. https://doi.org/10.1038/s41598-020-80149-9

Dewitte A, Twyford AD, Thomas DC, Kidner CA, Van Huylenbroeck J. 2011. The origin of diversity in *Begonia*: genome dynamism, population processes and phylogenetic patterns. In: Grillo O, Venora G, editors. The Dynamical Processes of Biodiversity – Case Studies of Evolution and Spatial Distribution. Rijecka, Croatia: InTech. pp. 27–52. https://cdn.intechopen.com/pdfs/24409.pdf

Dierckxsens N, Mardulyn P, Smits G. 2017. NOVOPlasty: *de novo* assembly of organelle genomes from whole genome data. Nucleic Acids Research. 45(4):e18. https://doi.org/10.1093/nar/gkw955

Dong LN, Du XY, Zhou W. 2019. The complete plastid genome sequence of *Begonia guangxiensis*. Mitochondrial DNA Part B-Resources. 4(2):3766–3767. https://doi.org/10.1080/23802359.2019.1681322

Downie SR, Jansen RK. 2015. A comparative analysis of whole plastid genomes from the Apiales: expansion and contraction of the inverted repeat, mitochondrial to plastid transfer of DNA, and identification of highly divergent noncoding regions. Systematic Botany. 40(1):336–351. https://doi.org/10.1600/036364415X686620

Doyle JJ, Davis JI, Soreng RJ, Garvin D, Anderson MJ. 1992. Chloroplast DNA inversions and the origin of the grass family (Poaceae). Proceedings of the National Academy of Sciences of the United States of America. 89(16):7722–7726. https://doi.org/10.1073/pnas.89.16.7722

Ebert D, Peakall R. 2009. Chloroplast simple sequence repeats (cpSSRs): technical resources and recommendations for expanding cpSSR discovery and applications to a wide array of plant species. Molecular Ecology Resources. 9(3):673–690. https://doi.org/10.1111/j.1755-0998.2008.02319.x

Fan J, Li XJ, Li CH, Yan BN. 2019. The complete chloroplast genome and phylogenetic analysis of *Begonia pulchrifolia*, a near endangered Begoniaceae plant. Mitochondrial DNA Part B-Resources. 4(2):2830–2831. https://doi.org/10.1080/23802359.2019.1660268

Gao C, Ren X, Mason AS, Liu H, Xiao M, Li J, Fu D. 2014. Horizontal gene transfer in plants. Functional and Integrative Genomics. 14:23–29. https://doi.org/10.1007/s10142-013-0345-0

Garrido-Ramos MA. 2015. Satellite DNA in plants: more than just rubbish. Cytogenetic and Genome Research. 146(2):153–170. https://doi.org/10.1159/000437008

Goodall-Copestake WP, Harris DJ, Hollingsworth PM. 2009. The origin of a mega-diverse genus: dating *Begonia* (Begoniaceae) using alternative datasets, calibrations and relaxed clock methods. Botanical Journal of the Linnean Society. 159(3):363–380. https://doi.org/10.1111/j.1095-8339.2009.00948.x

Goodall-Copestake WP, Perez-Espona S, Harris DJ, Hollingsworth PM. 2010. The early evolution of the mega-diverse genus *Begonia* (Begoniaceae) inferred from organelle DNA phylogenies. Biological Journal of the Linnean Society. 101(2):243–250. https://doi.org/10.1111/j.1095-8312.2010.01489.x

Goulding SE, Wolfe KH, Olmstead RG, Morden CW. 1996. Ebb and flow of the chloroplast inverted repeat. Molecular and General Genetics. 252:195–206. https://doi.org/10.1007/BF02173220

Gray MW. 1989. The evolutionary origins of organelles. Trends in Genetics. 5(9):294–299.

Green BR. 2011. Chloroplast genomes of photosynthetic eukaryotes. Plant Journal. 66(1):34–44. https://doi.org/10.1016/0168-9525(89)90111-X

Greiner S, Lehwark P, Bock R. 2019. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. Nucleic Acids Research. 47(W1):W59–W64. https://doi.org/10.1093/nar/gkz238

Guisinger MM, Kuehl JV, Boore JL, Jansen RK. 2011. Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and codon usage. Molecular Biology and Evolution. 28(1):583–600. https://doi.org/10.1093/molbev/msq229

Gymrek M, Willems T, Guilmatre A, Zeng HY, Markus B, Georgiev S, Daly MJ, Price AL, Pritchard JK, Sharp AJ, Erlich Y. 2016. Abundant contribution of short tandem repeats to gene expression variation in humans. Nature Genetics. 48:22–29. https://doi.org/10.1038/ng.3461

Harrison N, Harrison RJ, Kidner CA. 2016. Comparative analysis of *Begonia* plastid genomes and their utility for species-level phylogenetics. PLoS One. 11(4):0153248. https://doi.org/10.1371/journal.pone.0153248

Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, Bank van der M, Chase MW, Cowan RS, Erickson DL, Fazekas AJ, *et al.*; CBOL Plant Working Group. 2009. A DNA barcode for land plants. Proceedings of the National Academy of Sciences of the United States of America. 106(31):12794–12797. https://doi.org/10.1111/1755-0998.12194

Huang LP, Wang JM. 2020. Characterization of the complete chloroplast genome of *Begonia* *fimbristipula* (Begoniaceae). Mitochondrial DNA Part B-Resources. 5(1):774–775. https://doi.org/10.1080/23802359.2020.1715872

Hughes M, Girmansyah D. 2011. A revision of *Begonia* sect. *Sphenanthera* (Hassk.) Warb. from Sumatra. Gardens' Bulletin Singapore. 62(2):27–39. https://www.rbge.org.uk/media/3748/hughes-girmansyah-2011-a-revision-of-begonia-sect-sphenanthera-hassk-warb-from-sumatra.pdf

Hughes M, Hollingsworth PM, Miller AG. 2003. Population genetic structure in the endemic *Begonia* of the Socotra archipelago. Biological Conservation. 113(2):277–284. https://doi.org/10.1016/S0006-3207(02)00375-0

Hughes M, Rubite RR, Blanc P, Chung KF, Peng CI. 2015. The Miocene to Pleistocene colonization of the Philippine Archipelago by *Begonia* sect. *Baryandra* (Begoniaceae). American Journal of Botany. 102(5):695–706. https://doi.org/10.3732/ajb.1400428

Hughes M, Moonlight PW, Jara-Muñoz A, Tebbitt MC, Wilson HP, Pullan M. 2015–. Begonia Resource Centre. Online database. https://padme.rbge.org.uk/begonia/

Hughes M, Peng CI, Lin CW, Rubite RR, Blanc P, Chung KF. 2018. Chloroplast and nuclear DNA exchanges among *Begonia* sect. *Baryandra* species (Begoniaceae) from Palawan Island, Philippines, and descriptions of five new species. PLoS One. 13(5):0194877. https://doi.org/10.1371/journal.pone.0194877

Jansen RK, Ruhlman TA. 2012. Plastid genomes of seed plants. In: Bock R, Knoop V, editors. Genomics of Chloroplasts and Mitochondria. Dordrecht: Springer. pp. 103–126.

Jernigan KK, Bordenstein SR. 2015. Tandem-repeat protein domains across the tree of life. PeerJ. 3:732. https://doi.org/10.7717/peerj.732

Jin JJ, Yu WB, Yang JB, Song Y, Depamphilis CW, Yi TS, Li DZ. 2020. GetOrganelle: a fast and versatile toolkit for accurate *de novo* assembly of organelle genomes. Genome Biology. 21(1):241. https://doi.org/10.1186/s13059-020-02154-5

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution. 30(4):772–780. https://doi.org/10.1093/molbev/mst010

Khan A, Asaf S, Khan AL, Al-Harrasi A, Al-Sudairy O, Abdulkareem NM, Khan A, Shehzad T, Alsaady N, Al-Lawati A, Al-Rawahi A, Shinwari ZK. 2019. First complete chloroplast genomics and comparative phylogenetic analysis of *Commiphora gileadensis* and *C. foliacea*: myrrh producing trees. PLoS One. 14(1):0208511. https://doi.org/10.1371/journal.pone.0208511

Kim YD, Jansen RK. 1994. Characterization and phylogenetic distribution of a chloroplast DNA rearrangement in the Berberidaceae. Plant Systematics and Evolution. 193:107–114.

Kleffmann T, Russenberger D, von Zychlinski A, Christopher W, Sjölander K, Gruissem W, Baginsky S. 2004. The *Arabidopsis thaliana* chloroplast proteome reveals pathway abundance and novel protein functions. Current Biology. 14(5):354–362. https://doi.org/10.1016/j.cub.2004.02.039

Krishna N, Britto SJ, Thomas S, Mani B, Pradeep A, Jithin KV. 2020. A new section (*Begonia* sect. *Flocciferae* sect. nov.) and two new species in Begoniaceae from the Western Ghats of India. Edinburgh Jouranl of Botany. 77(2):251–268. https://doi.org/10.1017/S0960428619000349

Kurtz S, Choudhuri JV, 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. https://doi.org/10.1093/nar/29.22.4633

Lee HL, Jansen RK, Chumley TW, Kim KJ. 2007. Gene relocations within chloroplast genomes of *Jasminum* and *Menodora* (Oleaceae) are due to multiple, overlapping inversions. Molecular Biology and Evolution. 24(5):1161–1180. https://doi.org/10.1093/molbev/msm036

Li B, Zheng YQ. 2018. Dynamic evolution and phylogenomic analysis of the chloroplast genome in Schisandraceae. Scientific Reports. 8:9285. https://doi.org/10.1038/s41598-018-27453-7

Li ZH, Ma X, Wang DY, Li YX, Wang CW, Jin XH. 2019. Evolution of plastid genomes of *Holcoglossum* (Orchidaceae) with recent radiation. BMC Evolutionary Biology. 19:63. https://doi.org/10.1186/s12862-019-1384-5

Liu SH, Tseng YH, Zure D, Rubite RR, Balangcod TD, Peng CI, Chung KF. 2019. *Begonia balangcodiae* sp. nov. from northern Luzon, the Philippines and its natural hybrid with *B. crispipila*, *B. × kapangan nothosp. nov*. Phytotaxa. 407(1):5–12. https://doi.org/10.11646/phytotaxa.407.1.3

Maddison WP, Maddison DR. 2015. Mesquite: a modular system for evolutionary analysis. Version 3.5. http://www.mesquiteproject.org

Massouh A, Schubert J, Yaneva-Roder L, Ulbricht-Jones ES, Zupok A, Johnson MTJ, Wright SI, Pellizzer T, Sobanski J, Bock R, Greiner S. 2016. Spontaneous chloroplast mutants mostly occur by replication slippage and show a biased pattern in the plastome of *Oenothera*. The Plant Cell. 28(4):911–929. https://doi.org/10.1105/tpc.15.00879

McNeal JR, Kuehl JV, Boore JL, de Pamphilis CW. 2007. Complete plastid genome sequences suggest strong selection for retention of photosynthetic genes in the parasitic plant genus *Cuscuta*. BMC Plant Biology. 7:57. https://doi.org/10.1186/1471-2229-7-57

Moonlight PW, Ardi WH, Padilla LA, Chung KF, Fuller D, Girmansyah D, Hollands R, Jara-Muñoz A, Kiew R, Leong WC, Liu Y, Mahardika A, Marasinghe LDK, O'Connor M, Peng CI, Pérez AJ, Phutthai T, Pullan M, Rajbhandary S, Reynel C, Rubite RR, Sang J, Scherberich D, Shui YM, Tebbitt MC, Thomas DC, Wilson HP, Zaini NH, Hughes M. 2018. Dividing and conquering the fastest-growing genus: towards a natural sectional classification of the mega-diverse genus *Begonia* (Begoniaceae). Taxon. 67(2):267–323. https://doi.org/10.12705/672.3

Nock CJ, Waters DLE, Edwards MA, Bowen SG, Rice N, Cordeiro GM, Henry RJ. 2011. Chloroplast genome sequences from total DNA for plant identification. Plant Biotechnology Journal. 9(3):328–333. https://doi.org/10.1111/j.1467-7652.2010.00558.x

Palmer JD. 1987. Chloroplast DNA evolution and biosystematics uses of chloroplast DNA variation. The American Naturalist. 130:S6–S29. https://doi.org/10.1086/284689

Park S, An B, Park S. 2018. Reconfiguration of the plastid genome in *Lamprocapnos spectabilis*: IR boundary shifting, inversion, and intraspecific variation. Scientific Reports. 8:13568. https://doi.org/10.1038/s41598-018-31938-w

Picart-Picolo A, Grob S, Picault N, Franek M, Llauro C, Halter T, Maier TR, Jobet E, Descombin J, Zhang P, Paramasivan V, Baum TJ, Navarro L, Dvořáčková M, Mirouze M, Pontvianne F. 2020. Large tandem duplications affect gene expression, 3D organization, and plant–pathogen response. 30(11):1583–1592. https://doi.org/10.1101/gr.261586.120

Plana V. 2003. Phylogenetic relationships of the Afro-Malagasy members of the large genus *Begonia* inferred from *trnL* intron sequences. Systematic Botany. 28(4):693–704. http://www.jstor.org/stable/25063916

Plana V, Gascoigne A, Forrest LL, Harris D, Pennington RT. 2004. Pleistocene and pre-pleistocene *Begonia* speciation in Africa. Molecular Phylogenetics and Evolution. 31(2):449–461. https://doi.org/10.1016/j.ympev.2003.08.023

Qi WC, Lin F, Liu YH, Huang BQ, Cheng JH, Zhang W, Zhao H. 2016. High-throughput development of simple sequence repeat markers for genetic diversity research in *Crambe abyssinica*. BMC Plant Biology. 16:139. https://doi.org/10.1186/s12870-016-0828-y

Qu YJ, Legen J, Arndt J, Henkel S, Hoppe G, Thieme C, Ranzini G, Muino JM, Weihe A, Ohler U, Weber G, Ostersetzer O, Schmitz-Linneweber C. 2018. Ectopic transplastomic expression of a synthetic *matK* gene leads to cotyledon-specific leaf variegation. Frontiers in Plant Science. 9:1453. https://doi.org/10.3389/fpls.2018.01453

Raubeson LA, Peery R, Chumley TW, Dziubek C, Fourcade HM, Boore JL, Jansen RK. 2007. Comparative chloroplast genomics: analyses including new sequences from the angiosperms *Nuphar advena* and *Ranunculus macranthus*. BMC Genomics. 8:174. https://doi.org/10.1186/1471-2164-8-174

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic Biology. 61(3):539–542. https://doi.org/10.1093/sysbio/sys029

Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, Sánchez-Gracia A. 2017. DnaSP 6: DNA sequence polymorphism analysis of large data sets. Molecular Biology and Evolution. 34(12):3299–3302. https://doi.org/10.1093/molbev/msx248

Ruhlman TA, Jansen RK. 2014. The plastid genomes of flowering plants. In: Maliga P, editor. Chloroplast Biotechnology. Totowa, New Jersey, USA: Humana Press. pp. 3–38.

Ruhlman T, Lee SB, Jansen RK, Hostetler JB, Tallon LJ, Town CD, Daniell H. 2006. Complete plastid genome sequence of *Daucus carota*: implications for biotechnology and phylogeny of angiosperms. BMC Genomics. 7:222. https://doi.org/10.1186/1471-2164-7-222

Samson N, Bausher MG, Lee SB, Jansen RK, Daniell H. 2007. The complete nucleotide sequence of the coffee (*Coffea arabica* L.) chloroplast genome: organization and implications for biotechnology and phylogenetic relationships amongst angiosperms. Plant Biotechnology Journal. 5(2):339–353. https://doi.org/10.1111/j.1467-7652.2007.00245.x

Shen GF, Chen K, Wu M, Kung SD. 1982. *Nicotiana* chloroplast genome IV. *N. accuminata* has larger inverted repeats and genome size. Molecular and General Genetics. 187(1):12–18.

Sheue CR, Pao SH, Chien LF, Chesson P, Peng CI. 2012. Natural foliar variegation without costs? The case of *Begonia*. Annals of Botany. 109(6):1065–1074. https://doi.org/10.1093/aob/mcs025

Shui YM, Chen WH, Peng H, Huang SH, Liu ZW. 2019. Taxonomy of Begonias. Kunming: Yunnan Science and Technology Press.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30(9):1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Stefanović S, Rice DW, Palmer JD. 2004. Long branch attraction, taxon sampling, and the earliest angiosperms: *Amborella* or monocots? BMC Evolutionary Biology. 4:35. https://doi.org/10.1186/1471-2148-4-35

Straub SCK, Parks M, Weitemier K, Fishbein M, Cronn RC, Liston A. 2012. Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. American Journal of Botany. 99(2):349–364. https://doi.org/10.3732/ajb.1100335

Sun YX, Moore MJ, Meng AP, Soltis PS, Soltis DE, Li JQ, Wang HC. 2013. Complete plastid genome sequencing of Trochodendraceae reveals a significant expansion of the inverted repeat and suggests a Paleogene divergence between the two extant species. PLoS One. 8(4):e60429. https://doi.org/10.1371/journal.pone.0060429

Sun YX, Moore MJ, Zhang SJ, Soltis PS, Soltis DE, Zhao TT, Meng AP, Li XD, Li JQ, Wang HC. 2016. Phylogenomic and structural analyses of 18 complete plastomes across nearly all families of early-diverging eudicots, including an angiosperm-wide analysis of IR gene content evolution. Molecular Phylogenetics and Evolution. 96:93–101. https://doi.org/10.1016/j.ympev.2015.12.006

Tebbitt MC. 2005. *Begonia*: Cultivation, Identification, and Natural History. Portland, Oregon: Timber Press.

Thode VA, Lohmann LG. 2019. Comparative chloroplast genomics at low taxonomic levels: a case study using *Amphilophium* (Bignonieae, Bignoniaceae). Frontiers in Plant Science. 10:796. https://doi.org/10.3389/fpls.2019.00796

Thomas DC, Hughes M, Phutthai T, Rajbhandary S, Rubite R, Ardi WH, Richardson JE. 2011. A non-coding plastid DNA phylogeny of Asian *Begonia* (Begoniaceae): evidence for morphological homoplasy and sectional polyphyly. Molecular Phylogenetics and Evolution. 60(3):428–444. https://doi.org/10.1016/j.ympev.2011.05.006

Thomas DC, Hughes M, Phutthai T, Ardi WH, Rajbhandary S, Rubite R, Twyford AD, Richardson JE. 2012. West to east dispersal and subsequent rapid diversification of the mega-diverse genus *Begonia* (Begoniaceae) in the Malesian archipelago. Journal of Biogeography. 39(1):98–113. https://doi.org/10.1111/j.1365-2699.2011.02596.x

Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, Greiner S. 2017. GeSeq – versatile and accurate annotation of organelle genomes. Nucleic Acids Research. 45(W1):W6–W11. https://doi.org/10.1093/nar/gkx391

Tseng YH, Huang HY, Xu WB, Yang HA, Liu Y, Peng CI, Chung KF. 2017. Development and characterization of EST-SSR markers for *Begonia luzhaiensis* (Begoniaceae). Applications in Plant Sciences. 5(5):1700024. https://doi.org/10.3732/apps.1700024

Tseng YH, Huang HY, Xu WB, Yang HA, Peng CI, Liu Y, Chung KF. 2019. Phylogeography of *Begonia luzhaiensis* suggests both natural and anthropogenic causes for the marked population genetic structure. Botanical Studies. 60(1):e20. https://doi.org/10.1186/s40529-019-0267-9

Twyford AD, Ness RW. 2017. Strategies for complete plastid genome sequencing. Molecular Ecology Resources. 17(5):858–868. https://doi.org/10.1111/1755-0998.12626

Twyford AD, Ennos RA, Kidner CA. 2013a. Development and characterization of microsatellite markers for Central American *Begonia* sect. *Gireoudia* (Begoniaceae). Applications in Plant Sciences. 1(5):1200499. https://doi.org/10.3732/apps.1200499

Twyford AD, Kidner CA, Harrison N, Ennos RA. 2013b. Population history and seed dispersal in widespread Central American *Begonia* species (Begoniaceae) inferred from plastome-derived microsatellite markers. Botanical Journal of the Linnean Society. 171(1):260–276. https://doi.org/10.1111/j.1095-8339.2012.01265.x

Wang RJ, Cheng CL, Chang CC, Wu CL, Su TM, Chaw SM. 2008. Dynamics and evolution of the inverted repeat–large single copy junctions in the chloroplast genomes of monocots. BMC Evolutionary Biology. 8:36. https://doi.org/10.1186/1471-2148-8-36

Wang ZF, Liu TH, Cao HL. 2021. The complete chloroplast genome sequence of *Begonia coptidifolia*. Mitochondrial DNA Part B-Resources. 6(2):548–549. https://doi.org/10.1080/23802359.2021.1872434

Weng ML, Blazier JC, Govindu M, Jansen RK. 2014. Reconstruction of the ancestral plastid genome in Geraniaceae reveals a correlation between genome rearrangements, repeats, and nucleotide substitution rates. Molecular Biology and Evolution. 31(3):645–659. https://doi.org/10.1093/molbev/mst257

Weng ML, Ruhlman TA, Jansen RK. 2017. Expansion of inverted repeat does not decrease substitution rates in *Pelargonium* plastid genomes. New Phytologist. 214(2):842–851. https://doi.org/10.1111/nph.14375

Wolfe KH. 1991. Protein-coding genes in chloroplast DNA: compilation of nucleotide sequences, data base entries, and rates of molecular evolution. In: Bogorad L, Vasil IK, editors. The Photosynthetic Apparatus: Molecular Biology and Operation. Cell Culture and Somatic Cell Genetics of Plants, vol. 7B. New York: Academic Press. pp. 467–482.

Wu CY, Ku TC. 1997. New taxa of the *Begonia* L. (Begoniaceae) from China (cont.). Acta Phytotaxonomica Sinica. 35(1):43–56.

Xiong AS, Peng RH, Zhuang J, Gao F, Zhu B, Fu XY, Xue Y, Jin XF, Tian YS, Zhao W, Yao QH. 2009. Gene duplication, transfer, and evolution in the chloroplast genome. Biotechnology Advances. 27(4):340–347. https://doi.org/10.1016/j.biotechadv.2009.01.012

Xu C, Dong WP, Li WQ, Lu YZ, Xie XM, Jin XB, Shi JP, He KH, Suo ZL. 2017. Comparative analysis of six *Lagerstroemia* complete chloroplast genomes. Frontiers in Plant Science. 8:15. https://doi.org/10.3389/fpls.2017.00015

Zhang X, Sun YX, Landis JB, Lv ZY, Shen J, Zhang HJ, Lin N, Li LJ, Sun J, Deng T, Sun H, Wang HC. 2020. Plastome phylogenomic study of Gentianeae (Gentianaceae): widespread gene tree discordance and its association with evolutionary rate heterogeneity of plastid genes. BMC Plant Biology. 20:340. https://doi.org/10.1186/s12870-020-02518-w

Zhao XM, Su L, Schaack S, Sadd BM, Sun C. 2018. Tandem repeats contribute to coding sequence variation in Bumblebees (Hymenoptera: Apidae). Genome Biology and Evolution. 10(12):3176–3187. https://doi.org/10.1093/gbe/evy244

Zhou S, Wen M, Wang LY, Wang XY, Guo W, Xu YF. 2020. The complete chloroplast genome sequence of *Begonia versicolor* Irmsch. (Begoniaceae). Mitochondrial DNA Part B-Resources. 5(3):2113–2114. https://doi.org/10.1080/23802359.2020.1765706

Zhu AD, Guo WH, Gupta S, Fan WS, Mower JP. 2016. Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. New Phytologist. 209(4):1747–1756. https://doi.org/10.1111/nph.13743

Zoschke R, Nakamura M, Liere K, Sugiura M, Borner T, Schmitz-Linneweber C. 2010. An organellar maturase associates with multiple group II introns. Proceedings of the National Academy of Sciences of the United States of America. 107(7):3245–3250. https://doi.org/10.1073/pnas.0909400107

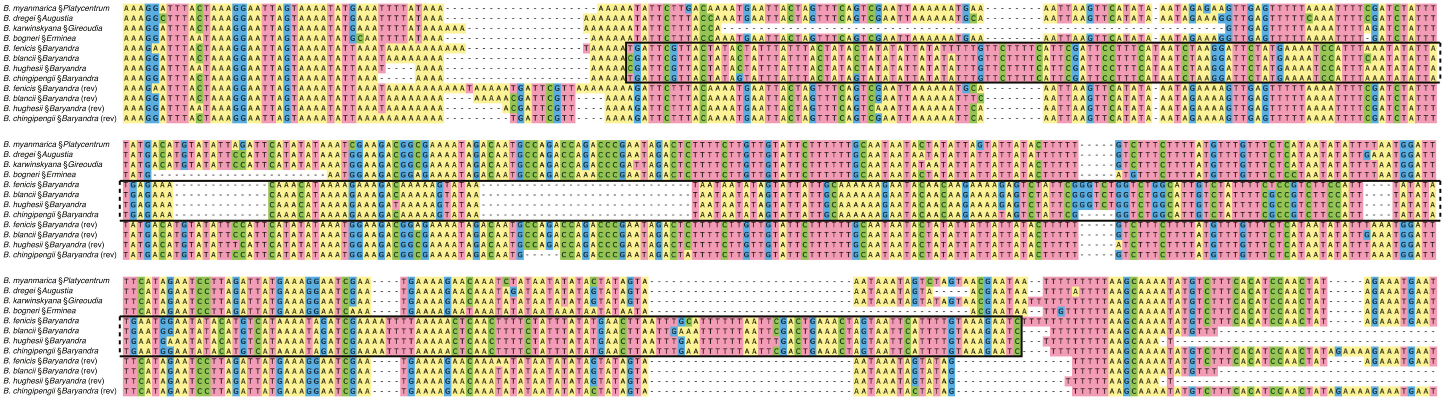
Appendix

Appendix table. Species used in the present study, with voucher and collection information

Species	Section	Voucher	Collection locality
Begonia			
B. aconitifolia A.DC.	*Latistigma*	*Peng* 22224	Brazil
B. albococcinea Hook.	*Flocciferae*	*Peng* 23302	India
B. amoeboides Moonlight	*Cyathocnemis*	*RBGE* 20180924	Peru: Amazonas Region, Prov. Bagua, road from Bagua to Rioja
B. ampla Hook.f.	*Squamibegonia*	*Peng* 22543	NA
B. anemoniflora Irmsch.	*Eupetalum*	*RBGE* 20160123	Peru: Junin Region, Prov. Concepcion, Dist. Comas.
B. anisosepala Hook.f.	*Scutobegonia*	*Peng* 24894	NA
B. baccata Hook.f.	*Baccabegonia*	*K* 023612	NA
B. bogneri Ziesenh.	*Erminea*	*Peng* 22541	Madagascar
B. bracteata Jack	*Bracteibegonia*	*Peng* 23521	Sumatra, Gunung Bungkuk
B. buddleiifolia A.DC.	*Pilderia*	*RBGE* 20160126	Peru: San Martin Region, Prov. San Martin, Tarapoto
B. chlorosticta Sands	*Petermannia*	*Peng* 23304	Malaysia: Borneo, Sarawak
B. convolvulacea (Klotzsch ex Klotzsch) A.DC.	*Wageneria*	*Peng* 21267	Brazil
B. cubensis Hassk.	*Begonia*	*Peng* 21285	Cuba
B. depauperata Schott	*Trachelocarpus*	*Peng* 24271	America
B. dipetala Graham	*Haagea*	*Peng* 22521	NA
B. dregei Otto & A.Dietr.	*Augustia*	*K* 19503	Africa
B. egregia N.E.Br.	*Tetrachia*	*Peng* 23327	NA
B. fenicis Merr.	*Baryandra*	*Peng* 10794	Taiwan: Lanyu, Mt Tasenshan
B. fissistyla Irmsch.	*Hydristyles*	*Peng* 21417	NA
B. foliosa Kunth	*Lepsia*	*Peng* 21254	NA
B. henrilaportei Scherber. & Duruiss.	*Nerviplacentaria*	*RBGE* 20160414	Madagascar
B. henryi Hemsl.	*Reichenheimia*	*RBGE* 20141517	China: Yunnan
B. heydei C.DC.	*Urniformia*	*RBGE* 20131992	Costa Rica
B. karwinskyana A.DC.	*Gireoudia*	*Peng* 20880	Mexico
B. kingiana Irmsch.	*Ridleyella*	*Peng* 21226	Malaysia
B. komoensis Irmsch.	*Tetraphila*	*Peng* 21211	NA
B. ludwigii Irmsch.	*Knesebeckia*	*Peng* 22333	Ecuador
B. meyeri-johannis Engl.	*Exalabegonia*	*RBGE* 20131229	Tanzania: Morogoro, Tchenzema, Ulruguru Nature Reserve
B. microsperma Warb.	*Loasibegonia*	*Peng* 20259	NA
B. myanmarica C.I Peng & Y.D.Kim	*Platycentrum*	*Peng* 23566	Myanmar: Sagaing region, Alangdaw Kathapa National Park

B. oaxacana A.DC.	*Parietoplacentalia*	*RBGKew* (*s.n.*)	Central America
B. oxyloba Welw. ex Hook.f.	*Exalabegonia*	*RBGE* 19982761	Tanzania: Amani Nature Reserve
B. picturata Yan Liu, S.M.Ku & C.I Peng	*Coelocentrum*	*Peng* 20387	China: Guangxi, Baise City, Jingxi County, Dizhou Township, Guwen Villlage
B. ravenii C.I Peng & Y.K.Chen	*Diploclinium*	*Peng* 22752	Taiwan: Nantou Hsien, Tsaotun Town, Shangtung
B. rossmanniae A.DC.	*Rossmannia*	*RBGE* 20151093	Peru: Pasco Region, Prov. Oxapampa, PN Yanachaga-Chemillen
B. samhaensis M.Hughes & A.G.Mill.	*Peltaugustia*	*RBGE* 19990398	Yemen: Socotra
B. sanguinea Raddi	*Pritzelia*	*Peng* 21284	Brazil
B. santos-limae Brade	*Astrothrix*	*Peng* 21320	NA
B. tigrina Kiew	*Jackia*	*Peng* 22720	Malaysia: Greenhouse of Forest Research Institute of Malaysia
B. ulmifolia Willd.	*Donaldia*	*RBGE* 20030607	Cultivated, of no known wild origin
B. undulata Schott	*Gaerdtia*	*Peng* 21275	NA
B. variabilis Ridl.	*Parvibegonia*	*Peng* 21040	Thailand: Nakhon Si Thammarat Province, Nop Pitum District
B. sp. nov., sect. *Ruizopavonia*	*Ruizopavonia*	*RBGE* 20160139	Peru: Ucayali Region, Coronel Portillo Province, Dist. Padre Abad, Boqueron de Padre Abad
Hillebrandia sandwicensis Oliv.		*Natalia Tangalin* 4564	Hawai‘i: Kokee

NA, not available.


B. myanmarica §Platycentrum
B. dregei §Augustia
B. karwinskyana §Gireoudia
B. bogneri §Erminea
B. fenicis §Baryandra
B. blancii §Baryandra
B. hughesii §Baryandra
B. chingipengii §Baryandra
B. fenicis §Baryandra (rev)
B. blancii §Baryandra (rev)
B. hughesii §Baryandra (rev)
B. chingipengii §Baryandra (rev)

Appendix igure. Alignment of the *ndhF–rpl32* sequence of *Begonia*. Five representative *Begonia* species from our study and three sect. *Baryandra* species from NCBI were aligned using MAFFT. Dashes indicate alignment gaps and missing data; the boxed area is a 213-bp inversion of the *ndhF–rpl32* spacer of sect. *Baryandra* species. This fragment was used to determine the reverse-complement to generate new combined sequences for comparison (denoted as 'rev' at the end of the sequence name). NCBI reference numbers used in this study: *Begonia blancii* M.Hughes, KR186537; *B. hughesii* Rubite & C.I Peng, KR186564; and *B. chingipengii* Rubite, KR186542.