國立臺灣大學生物資源暨農學院森林環境暨資源學系

碩士論文

School of Forestry and Resource Conservation

College of Bioresources and Agriculture

National Taiwan University

Master Thesis

中越石灰岩溶植物

－報春苣苔屬與石山苣苔屬之葉綠體基因體

Plastomes of *Primulina* and *Petrocodon*

－ Two Characteristic Plant Genera of Sino-Vietnamese Limestone Flora

謝佳倫

Chia-Lun Hsieh

指導教授：鍾國芳 博士

Advisor: Kuo-Fang Chung Ph.D.

中華民國 107 年 1 月

January 2018

國立臺灣大學碩士學位論文
口試委員會審定書

中越石灰岩溶植物
—報春苣苔屬與石山苣苔屬之葉綠體基因體

Plastomes of *Primulina* and *Petrocodon*
— Two Characteristic Plant Genera of Sino-Vietnamese Limestone Flora

本論文係謝佳倫君（r04625013）在國立臺灣大學森林環境暨資源學系、所完成之碩士學位論文，於民國 107 年 1 月 30 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

_____（簽名）

（指導教授）

_____　_____

_____

系主任、所長　_____（簽名）

i

謝辭

這篇論文的完成，必須感謝眾多師長與朋友的協助與鼓勵。首先要感謝我的指導教授鍾國芳老師，在我碩士班期間引導我、鼓勵我，並幫助我解決研究上的問題，和我討論研究的方向與論文的內容。感謝口試委員蘇惠君老師與劉少倫老師，願意撥冗審閱論文，提供寶貴建議與匡正。

感謝廣西植物研究所許為斌博士等人和鍾國芳老師的合作，奠基了此研究的材料收集，多次陪同或協助採集行程，並提供許多珍貴的植物材料。也感謝過去與鍾國芳老師一同前往廣西、湖南等華南各地採集的研究室學長姊、許翊萱和朱安雅。沒有大家的協助，勢必得花費極大的成本來獲得研究材料，謹致上由衷的謝意。另外，非常感謝黃雅怡博士於基因體組裝上的協助，也感謝翁茂倫博士耐心解惑我的疑問，使我組裝基因體的過程更加順利。

很感謝研究室學長姊與同學的協助與指導，一同度過歡樂的研究室生活與野外採集時光。非常感謝劉世慧學姊於分析方法上提供寶貴的建議與幫助，並仔細修改我的論文，使之更加完善；感謝黃群智學姊，帶領我熟悉分子技術相關實驗與實驗室的注意事項；感謝曾妤馨學姊，平時和我討論各方面的研究知識並提出許多精闢看法與建議，使我獲益良多；感謝游旨价學長，在研究之路上的督促與鼓勵；感謝郭聞喜學長，和我討論資料的分析方法和實驗上的技術指導；感謝黃郁嵐學長於研究上的建議，並總是邀我去野外看植物散心；很感謝許翊萱於實驗初期的協助，萃取出許多良好品質的 DNA 樣本，讓定序順利進行。感謝研究室的所有人在這段時間的陪伴與指教，都是讓我能夠完成論文的莫大推手。

最後，我要感謝我的父母，雖然每每在野外時總是讓他們掛心，非常感謝他們一直以來對我的包容與支持，最後我才能順利完成此論文研究。

Abstract

Primulina and *Petrocodon*, two characteristic Gesneriaceae genera of the limestone flora in southern China and northern Vietnam, are featured by their exclusive habitat preference and remarkable diversity. The genome skimming strategy, which sequences genomic DNA in low coverage to recover the high copy portions and to increase the multiplex level of sequencing, was employed to obtain complete plastomes, mitochondrial exomes and nuclear ribosomal DNAs of eight *Primulina* and two *Petrocodon* species. Phylogenies based on the data from the three cell compartments were reconstructed and compared. Additional eight Gesneriad plastomes (including five *Primulina* spp.) accessed online were incorporated in the comparative and phylogenetic analyses. The newly sequenced plastomes are highly conserved in genome structures and gene contents. Although the phylogenies inferred from plastid, mitochondrial and nuclear data are incongruent, the well supported plastome phylogeny represents a backbone relationships of *Primulina* and its allies. The sequence divergence among plastomes of *Primulina* and *Petrocodon* were evaluated to identify variable regions which serve as potentially useful markers for systematic or population studies.

Key words: Genome skimming, Karst flora, Mitochondrial exome, *Petrocodon*, Plastome, *Primulina*.

摘要

分布於中國華南與北越石灰岩地區之苦苣苔科(Gesneriaceae)報春苣苔屬 (*Primulina*)與石山苣苔屬(*Petrocodon*)，擁有特殊的生境偏好與豐富的多樣性，為當地喀斯特物種之代表類群。本研究採用基因體略讀(genome skimming)之定序策略，定序八個報春苣苔屬與兩個石山苣苔屬物種之低覆蓋率(coverage)全基因體資料，獲取在細胞中擁有高副本數(copy number)的葉綠體基因(plastome)、粒線體外顯子體(mitochondria exome)與細胞核核醣體 DNA(nuclear ribosomal DNA)之序列，並分別進行親緣關係樹的重建。此法不需透過成本較高的高覆蓋率定序，仍可獲得大量資料，並能夠增加同時定序的樣本數量。此外，八個下載自基因銀行 (GenBank)的苦苣苔科物種之完整葉綠體基因體亦被納入基因體比較與親緣關係分析。雖然不同胞器 DNA 所重建之親緣關係樹存在許多不一致之處，但根據完整葉綠體基因體重建之演化關係具有極高支持度，可作為報春苣苔屬與其近緣屬間演化歷史之骨幹架構。藉由加入其他物種之序列進行比較，可得知所有納入分析之葉綠體基因體於基因組成與基因體的結構上均相當保守。本研究亦藉由排列比對報春苣苔屬與石山苣苔屬之完整葉綠體基因體序列，辨識出數個高變異的基因體區段，可作為未來進行族群遺傳與親緣關係研究所需的高解析力分子標記。
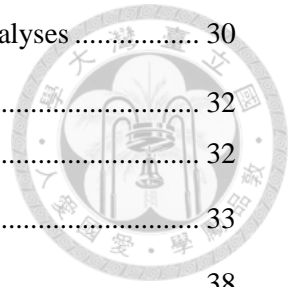

關鍵字: 報春苣苔屬、石山苣苔屬、葉綠體基因體、粒線體外顯子體、基因體略讀、喀斯特植物

Content

List of Figure

List of Table

1. Introduction

1.1 Gesneriaceae in karst landscape of southern China

Gesneriaceae Rich. & Juss. ex. DC., which comprises 150 – 160 genera and 3000 – 3500 species, is mainly distributed in the tropics and subtropics across the world (Weber, 2004; Skog and Boggan, 2006). In China, 41 genera and over 580 species have been recorded with, surprisingly, over 80% of endemism (Möller *et al.*, 2016). Most of them are distributed in southern China, where is covered by extensive limestone landscape (or so called Karst topography) (Xu, 1995). Therefore, southern China was recognized as a hotspot for Chinese Gesneriaceae (Wei *et al.*, 2004; Hou *et al.*, 2010).

The karst environment plays an important role in the diversity of southern China flora (Xu, 1995; Hou *et al.*, 2010) due to the complex and unique habitats it provides. High salinity, high porosity and fragmented habitats are the general features of limestone karst, making the landscape unsuitable for ordinary plants to grow. However, many of the Gesneriaceae in southern China prefer, or even grow exclusively on the harsh environment of limestone bedrock or acidic soil type (Wei *et al.*, 2004). The niche specificity of karst plants, along with the continual fragmentation of karst terrains, may facilitate allopatric speciation and isolation by distance, driving the speciation of those karst specialized plants, resulting the diverse limestone flora of southern China (Chang, 2012).

1

1.2 *Primulina* and *Petrocodon*

1.2.1 *Primulina* Hance

The genus *Primulina* Hance was established by H. F. Hance in 1883 as a monotypic genus containing only *P. tabacum* (Hance, 1883). Based on the molecular data, the genus was expanded to include *Chirita* Buch.-Ham. ex. D. Don section *Gibbosaccus* C. B. Clarke, *Chiritopsis* W. T. Wang and *Wentsaiboea* D. Fang & D. H. Qin (Wang *et al.*, 2011; Weber *et al.*, 2011a). Before the revisions, the genus *Chirita* was morphologically diverse containing over 100 species with two-thirds of which belonged to the sect. *Gibbosaccus*, but has long been problematic in the generic and infra-generic delimitation (Li and Wang, 2007; Möller *et al.*, 2009). The species number of sect. *Gibbosaccus* increased rapidly due to many new species found in southern China karst area. After the re-circumscription of *Chirita* and *Primulina*, the latter has become one of the largest genus of Old World Gesneriaceae largely owing to the inclusion of the species-rich *Chirita* sect. *Gibbosaccus* (Wang *et al.*, 2011; Weber *et al.*, 2011a). As many new species of *Primulina* that were being described until recent years, the genus is constantly increasing in species number (e.g., Yang *et al.*, 2016; Wu *et al.*, 2017; Yang *et al.*, 2017; Tian *et al.*, 2018). *Primulina* now contains over 170 predominantly calciphilous species growing almost exclusively on limestone area of southern China and few in northern Vietnam (Möller *et al.*, 2016). These two regions

2

are collectively referred to as Sino-Vietnamese Karst, which is characterized by ubiquitous karst landscape and diverse flora (Xu, 1995).

The rich diversity, high endemism and strict habitat preference have made *Primulina* suitable materials for studying species divergence, adaptation and speciation of limestone flora. However, many described species differ only among trivial morphological traits and some of them may be taxonomic synonyms (Christie *et al.*, 2012) The diversity of the genus was probably overestimated (Weber *et al.*, 2011a). Recent studies showed that adopting both morphological and molecular data have facilitated the specific and generic delimitation in *Primulina* (e.g., Xu *et al.*, 2012; Chung *et al.*, 2013; Pan *et al.*, 2013; Xu *et al.*, 2013). However, the evolutionary history of *Primulina* was not fully resolved due to the large species number and insufficient variability of the molecular makers (e.g., Kang *et al.*, 2014; Guo *et al.*, 2015). More informative molecular markers or sequence information are required to improve the resolution of the relationships among taxa and to answer more evolutionary questions of karst flora.

1.2.2 *Petrocodon* **Hance**
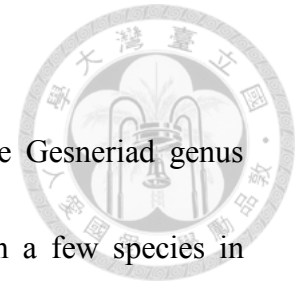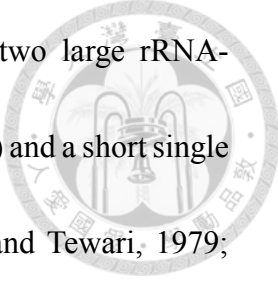
Petrocodon Hance is another calciphilous yet small Chinese Gesneriad genus which is distributed in Karst topography of southern China with a few species in northern Indochina (Hance, 1883; Chen *et al.*, 2014). The genus has been expanded by including the monotypic genera *Calcareoboea* C. Y. Wu ex. H. W. Li, *Dolicholoma* D. Fang & W. T. Wang, *Paralagarosolen* Y. G. Wei, and *Tengia* Chun, plus all species from *Lagarosolen* W. T. Wang, three species from *Didymocarpus* Wall., one species from *Wentsaiboea*, and one species from *Primulina* based on earlier molecular phylogenies (Möller *et al.*, 2011; Wang *et al.*, 2011; Weber *et al.*, 2011a; Weber *et al.*, 2011b; Xu *et al.*, 2014). The molecular data support that these genera formed a well-supported monophyletic group and is sister to *Primulina* (Möller *et al.*, 2011; Wang *et al.*, 2011; Weber *et al.*, 2011a; Weber *et al.*, 2011b; Xu *et al.*, 2014). Regarding to the various genera it had accommodated after the re-circumscription, *Petrocodon* now comprises about 30 species with extreme morphological diversity, especially in floral characteristics (Weber *et al.*, 2011b; Wen *et al.*, 2012; Xu *et al.*, 2014; Möller *et al.*, 2016; Cen *et al.*, 2017; Lu *et al.*, 2017). Weber *et al.* (2011b) considered the amazing variations in floral morphologies the consequence of rapid radiation evolving into various pollination syndromes. Therefore, *Petrocodon* is particularly ideal for studying floral evolutionary changes.

1.3 General features of plastome
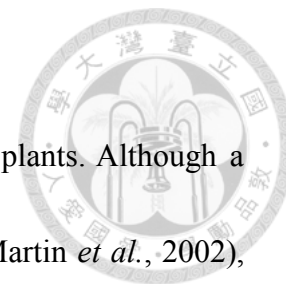
The chloroplast (cp) is a vital organelle for the survival of plants. Although a majority of the cp genes had transferred to the nuclear genome (Martin *et al.*, 2002), the endosymbiotic origin makes the plastid retains its own genome with multiple copies per organelle and be able to replicate by itself (Gray and Archibald, 2012). The remaining genes encoded in land plant plastomes are mainly responsible for photosynthesis and housekeeping functions, such as transcription and translation (Bock, 2007; Wicke *et al.*, 2011). Most of the land plant plastomes contain about 110 – 130 genes including 8 rRNA genes (4 duplicated genes in IRs), 30 – 35 tRNA genes and about 70 – 88 protein coding genes depending on plant lineages (Wicke and Schneeweiss, 2015). The gene content and the coding sequences are quite conserved across various lineages of land plants (Palmer and Stein, 1986; Wolf, 2012) because of the strong selective pressure acting on the genes, and the genes are crucial to the survival of plants (Rochaix, 1997).

The genome size of most of the land plant plastomes is more or less than 150 kb while ranging from 107 kb (*Cathaya argyrophylla*; Lin *et al.*, 2010) to 218 kb (*Pelarogium*; Chumley *et al.*, 2006), and it is not correlated with the nuclear genome size (Daniell *et al.*, 2016). The structure and gene orders of plastome is highly conserved among land plants as well (Palmer and Stein, 1986). The typical image of a

5

plastome is represented as a circular chromosome consisting of two large rRNA-encoding inverted repeats (IRs) separated by a long single copy (LSC) and a short single copy (SSC) regions, forming a quadripartite structure (Kolodner and Tewari, 1979; Palmer, 1985). The presence of two copies of IR enables the flip-flop recombination which results in isoforms of different orientations of the single copy regions (Palmer, 1983; Martin *et al.*, 2013). However, some studies have shown that the shapes of plastid DNAs are not necessarily circular with many of them found to be concatenated linear or branched, and probably vary among different lineages or stages of plastid development (Lilly *et al.*, 2001; Bendich, 2004; Oldenburg and Bendich, 2004; Shaver *et al.*, 2008). To be consistent with general publications of plastomes, the plastomes newly sequenced in this study were mapped to a circular conformation.

The inheritance of angiosperm plastids has long been thought to be predominantly maternal (Corriveau and Coleman, 1988; Birky, 1995; Mogensen, 1996; Zhang *et al.*, 2003; Hagemann, 2004). However, accumulating evidence has shown that some lineages exhibit bi-parental (e.g., Meyer and Stubbe, 1974; Metzlaff *et al.*, 1981; Nagata *et al.*, 1999) or even paternal inheritance (*Actinidia*; Testolin and Cipriani, 1997). The uniparental transmission of organelle is probably evolutionarily unstable (Greiner *et al.*, 2015).

1.4 The application of cpDNA to plant systematics
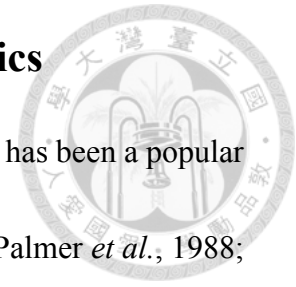
In plant molecular systematics, the chloroplast DNA (cpDNA) has been a popular and frequently adopted source of markers in the past few decades (Palmer *et al.*, 1988; Soltis and Soltis, 1998; Shaw *et al.*, 2005; Shaw *et al.*, 2014), even though the cpDNA evolves about half of the speed of nuclear DNA (Wolfe *et al.*, 1987). Despite the conserve nature, other characteristics of cpDNA, such as multiple copies, uniparental inheritance and the relatively small, haploid genome, make it an easily accessible data source representing a single genealogical history, which is convenient to phylogenetic analyses (Palmer *et al.*, 1988; Kelchner, 2000). These features of cpDNA can avoid the problems caused by recombination, polyploidy, or paralogues when using nuclear data (Feliner and Rosselló, 2007). Moreover, by comparing the phylogenetic histories inferred from cpDNA with nuclear phylogenies, hybridization hypotheses could be tested (e.g., Sang *et al.*, 1997; Isoda *et al.*, 2000; Barber *et al.*, 2007). With respect to highly conserved coding parts of plastome, intergenic regions have been employed and were proved to be useful to resolve phylogenies at lower taxonomic level (e.g., Shaw *et al.*, 2005; Shaw *et al.*, 2007; Shaw *et al.*, 2014). However, the tiny fractions of cpDNA amplified by PCR were not informative enough to resolve the relationships of many taxon-rich, radiated, or phylogenetically recalcitrant groups (e.g., Knox and Palmer, 1995; Turner *et al.*, 2013). Applying larger numbers of molecular markers or

7

longer sequences have been suggested to improve the resolution for the phylogenetic studies (Parks *et al.*, 2009).
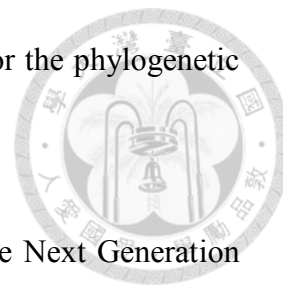
The advances in the sequencing technology, that is to say, the Next Generation Sequencing (NGS), in the past decade have made genome-scale data more available and affordable (Niedringhaus *et al.*, 2011). As a result, there have been over 1300 complete plastome sequences deposited in GenBank until 2016 (Tonti-Filippini *et al.*, 2017). Many studies have utilized the whole or partial plastome sequences to conduct phylogenomic analyses, yielding highly resolved phylogenetic trees (Parks *et al.*, 2009). This has been proven effective to resolve phylogenetic relationship of various taxonomic level ranging from intraspecific level (Parks *et al.*, 2009) to deep phylogeny among entire green plants (Ruhfel *et al.*, 2014).

The soaring number of plastid genomes being sequenced has also allowed a wide range of plant plastomes to be compared, facilitating our understanding of plastome evolution such as the seldom observed structural rearrangements or gene loss events (e.g., Ku *et al.*, 2013; Lei *et al.*, 2016; Song *et al.*, 2017). It is the conserved nature of plastid structure that makes inversions or gene/intron losses powerful characters to build phylogeny. For instance, Graham *et al.* (2000) have utilized the characters of indels and inversions in plastomes to decipher the deep phylogenetic relationship of basal angiosperms.

1.5 Next generation sequencing and genome skimming
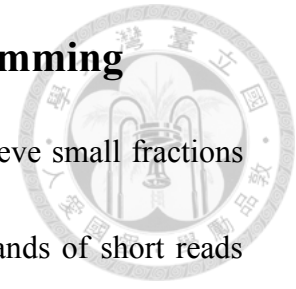
While traditional methods like Sanger sequencing only retrieve small fractions of DNA from the vast genome, NGS can sequence tens of thousands of short reads simultaneously, generating enormous molecular data in a short period of time (Niedringhaus *et al.*, 2011). To overcome the sequencing errors inherent to high throughput sequencers and to compensate for the shorter read length comparing to traditional PCR data, the same genomic site must be sequenced multiple times. This is the concept called 'coverage', or 'sequencing depth' of the NGS data (El-Metwally *et al.*, 2014). It is the short but numerous reads assembled into an intact molecule with enough coverage that the correct and complete genome is recovered (El-Metwally *et al.*, 2014). However, generating a complete nuclear genome with enough coverage requires expensive and laborious efforts, especially for the colossal genome of most of the plants. Moreover, since the sequence information of every single instrumental run of the NGS output is inevitably immense, it still costs quite a lot even though the cost of sequencing per base has dropped dramatically in recent years (Niedringhaus *et al.*, 2011). How to minimize the cost and maximize the value of NGS techniques is an important concern? To be cost-effective, several approaches of genome reduction methods, such as target enrichment and RAD-seq, have been developed just to sequence parts of the genome (Cronn *et al.*, 2012) by modifying the library preparation protocols

9

before the sequencing reactions.

An alternative way is to sequence a complete set of cellular DNA (Fig. 1) with low coverage, a method which may not be sufficient to recover the complete nuclear genome, but is desirable for those high copy portions in the cell, such as organelle genomes (plastids and mitochondria) and nuclear ribosomal DNAs (nrDNAs). This method which is named genome skimming (Straub *et al.*, 2012; Fig. 1), or genome surveying sequencing (Steele *et al.*, 2012) is amenable to multiplexing several samples in a single lane, raising the efficiency and reducing the cost of each run. Comparing to genome reduction methods, the library preparation steps of genome skimming are relatively simple for processing only the total cellular DNA without any enzymatic treatment or sequence baiting (Straub *et al.*, 2012). The method is particularly suitable for acquiring the phylogenetically useful plastome data, and has been applied extensively to the studies working on the comparative genomics and phylogenomic analyses using cpDNA (e.g., Malé *et al.*, 2014; Zhang *et al.*, 2015). Not only can genome skimming obtain plastome, but also it can recover the high copy nrDNAs and partial mitochondrial genome which may provide different genealogical insights to be compared (Henriquez *et al.*, 2014).

10
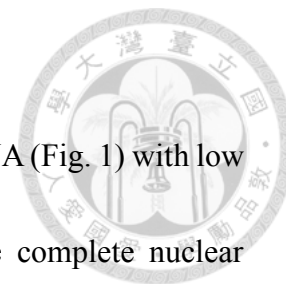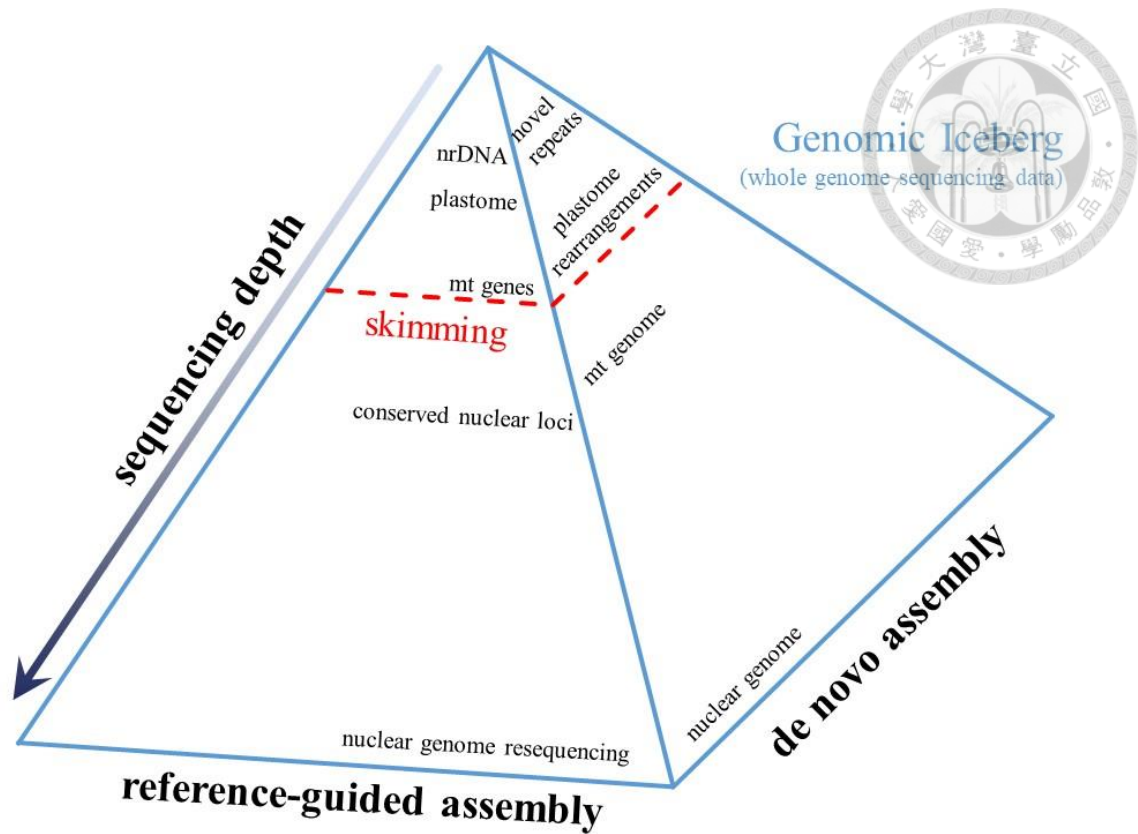
Fig. 1. The genomic iceberg explaining the concept of genome skimming.

The whole genome sequencing data is analogized as a genomic iceberg. To obtain the data in lower part of the iceberg, deeper sequencing depth is required. To reduce cost and increase efficiency, the genome skimming method only 'skim' the uppermost part (nrDNA, plastome and mt genes) of the iceberg, as indicated by the red dashed line. The figure is modified from Straub *et al.* (2012).

1.6 Goals of this study

Primulina and *Petrocodon*, being two characteristic genera of karst regions, can serve as decent materials to conduct evolutionary studies of limestone plants, probing into the underlying processes and mechanisms of plant evolution in this specific habitat. To resolve their phylogenetic relationships more robustly, I explored the complete plastome as an informative data source by using the genome skimming approach to reduce cost and increase sampling. The mitochondrial data, which is unpopular to systematics, and the nuclear ribosomal DNA can be acquired by this sequencing strategy as well to serve as different genealogical histories and to be compared with plastid data. Therefore, the study is aimed to:

(1) Sequence and assemble the complete plastomes of two limestone characteristic genera, *Primulina* and *Petrocodon*, by genome skimming methodology;

(2) Retrieve sequences from other cell compartments, mitochondria and nuclear (for nrDNA), to infer phylogenies and compare them to plastome data;

(3) Reconstruct the phylogenetic relationship using the newly sequenced and other published Gesneriaceae plastomes to provide a backbone relationship of the family;

(4) Identify variable regions among *Primulina* and related genera based on the aligned plastomes as applicable markers to better resolve the genus in the future.

2. Materials and Methods

2.1 Plant materials and taxon sampling

The plant materials were collected from the field expedition in Guangxi and Hunan, as well as the Guangxi Botanical Garden in 2015. The voucher specimens were deposited at Herbarium of Biodiversity Research Center, Academia Sinica, Taipei (HAST). Fresh leaves were dried by silica gel to preserve the DNA contents.

The taxa sequenced in this study were chosen as diverged as possible to represent each clades in the phylogenies of *Primulina* referring to the results of Guo *et al.* (2015). *Petrocodon* was also included as a sister group to *Primulina*. Regarding to the concentration and quality of extracted DNA, 11 samples with nine *Primulina* and two *Petrocodon* species were sent to the library preparation steps (Table 1).

In addition, the complete plastomes of five *Primulina* species (*P. brachytricha* var. *magnibracteata* (W. T. Wang & D. Y. Chen) Mich. Möller & A. Weber, *P. eburnea* (Hance) Y. Z. Wang, *P. huaijiensis* Z. L. Ning & J. Wang*, P. liboensis* (W. T. Wang & D. Y. Chen) Mich. Möller & A. Weber, *P. linearifolia* (W. T. Wang) Y. Z. Wang) and three other Gesneriad species (*Boea hygrometrica* (Bunge) R.Br., *Lysionotus pauciflorus* Maxim, and *Haberlea rhodopensis* Friv.) (Table 1 and Table 3), which are accessible in the GenBank to date, were also included in the comparative genomic analyses.

13
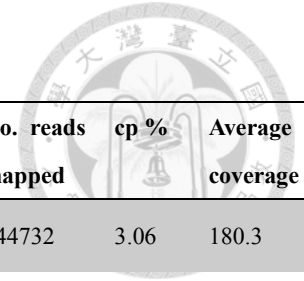
Table 1. Statistics of the assembly in the newly sequenced taxa.

Scientific name	Chinese name	Voucher No.	No. raw reads	Data size (Gb)	No. trimmed reads	No. reads mapped	cp %	Average coverage
Primulina lutea (Y. Liu & Y.G. Wei) Mich. Möller & A. Weber	黃花牛耳朵	KFC2916	7327898	1.8	4725427	144732	3.06	180.3
Primulina pengii W.B. Xu & K.F. Chung,	彭氏報春苣苔	Peng24024	8243930	2.1	5577411	175136	3.14	227.3
Primulina longgangensis (W.T. Wang) Y.Z. Wang	弄崗唇柱苣苔	Peng22948	7934422	2	5237240	59922	1.14	77.5
Primulina sclerophylla (W.T. Wang) Y.Z. Wang	硬葉唇柱苣苔	KFC2971	3196114	0.8	2211064	85092	3.85	111.8
Primulina hezhouensis (W.H. Wu & W.B. Xu) W.B. Xu & K.F. Chung	賀州小花苣苔	KFC2914	3196076	0.8	2494103	89047	3.57	119.1
Primulina medica (D. Fang) Y.Z. Wang	藥用唇柱苣苔	KFC2962	2784846	0.7	2088503	113210	5.42	150.9
Primulina fimbrisepala (Hand.-Mazz.) Y.Z. Wang	螞蝗七	KFC4144	2910610	0.7	2092014	78918	3.77	104.4
Primulina cordata (W.T. Wang) Mich. Möller & A. Weber	心葉唇柱苣苔	KFC3021	3180456	0.8	2182271	57091	2.62	74.7
Primulina lungzhouensis (W.T. Wang) Mich. Möller & A. Weber	龍州唇柱苣苔	KFC4184	988	0	-	-	-	-
Petrocodon coriaceifolius (Y.G. Wei) Y. G. Wei & Mich. Möller	革葉石山苣苔	KFC2943	2808524	0.7	2111802	132118	6.26	172.8
Petrocodon multiflorus F. Wen & Y.S. Jiang	多花石山苣苔	KFC2913	6254490	1.6	4377605	358083	8.18	435

Abbreviation: cp: chloroplast.

14

2.2 DNA extraction and next generation sequencing
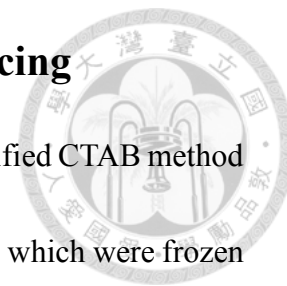
The total genomic DNA of each samples was extracted by modified CTAB method (Doyle and Doyle, 1987, 1990) from 0.01 g of silica gel dried leaves which were frozen by liquid nitrogen and ground into powder. The quantity and quality of extracted DNA were checked by 1% gel electrophoresis, Nanodrop 2000 (Thermo Scientific Inc., Carlsbad, CA, USA) and Qubit Fluorometer (Thermo Scientific).

The DNA samples passing the quality checks were sent to High Throughput Genomics Core at Biodiversity Research Center, Academia Sinica (BRCAS) to conduct library preparation and next generation sequencing. The DNA library were constructed by using TruSeq Nano DNA Library Prep kit (Illumina Inc., San Diego, CA, USA) following manufacturer's instruction with insert size 550 bp in length. Indexed DNA libraries were sequenced by Illumina MiSeq System (Illumina) in pair-end mode with read length = 300 bp. In total of 11 of my samples plus five samples of other species for other research projects were multiplexed in one flow cell of the sequencing machine which contains only lane.
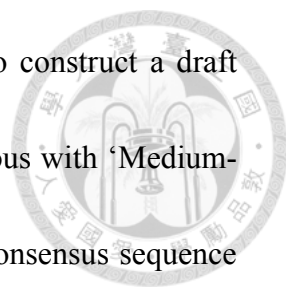
2.3 Plastome assembly and validation
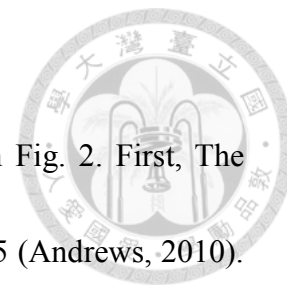
A brief work flow of the plastome assembly is presented in Fig. 2. First, The sequencing quality of raw reads were evaluated via FastQC v0.11.5 (Andrews, 2010). Low quality portions of the reads were trimmed and filtered by Trimmomatic 0.36 (Bolger *et al.*, 2014) with the setting of LEADING:25 TRAILING:25 SLIDINGWINDOW:4:20 MINLEN:200. The trimmed reads were then *de novo* assembled by Velvet assembler (Zerbino and Birney, 2008) with various combination of settings (kmer=141−181 with interval= 4; -cov_cutoff 50/100/200; -min_contig_lgth 100; -scaffolding yes/no), generating multiple contig files for each samples. Since the assembly is primarily purposed to recover the plastid DNA, higher coverage cutoff value will leave more contigs belonging to plastomes, which have multiple copies per cell. The number of contigs and N50 value of each combinations of different settings were compared to each other. For each samples, 6 − 10 contig files which have higher N50 value, fewer contigs number and longer overall contig length were chosen and imported into Geneious R10 (Kearse *et al.*, 2012) for subsequent analyses.

Second, the imported contigs and trimmed reads files were filter by Basic Local Alignment Tool (BLAST) in Geneious against the reference *B. hygrometrica* cp (Bhcp) genome (Zhang *et al.*, 2012) with Megablast mode and Max E-value= 10^{-5} to separate chloroplast and non-chloroplast reads or contigs. All the filtered cp contigs files of the

16

same species were then mapped to the reference Bhcp genome to construct a draft

plastome sequence by using 'Map to Reference' function in Geneious with 'Medium-

Low Sensitivity / Fast' and 'Iterate up to 25 times' settings. The consensus sequence

representing draft plastome of each species was generated with 50% masking threshold,

and a question mark masked areas with coverage under 1×. Next, the trimmed cp reads

were mapped to the draft genome by using 'Map to Reference' function and consensus

sequence was generated in Geneious with the same settings as the previous step to fill

the gaps. To validate the gap-filled genome, the trimmed cp reads were mapped back to

the latest consensus again. By screening the read mapping condition across each sites

of the plastome, I were able to find out and mark the regions with problematic

assemblies. The structural differences, which are usually large insertions and deletions

in length of 50 – 300 bp or longer that cannot be detected by read mapping method,

between reference and my species may cause the reads mapped on the genome forming

a 'chimeric alignment' (Fig. 3). To solve this problem, the draft sequence were cut into

pieces and were mapped by reads of the same species to extend the contigs. The draft

sequences were cut at flanking sites upstream and downstream to the problematic

regions to avoid extracting incorrect sequences being extracted. The cut parts were

served as references being mapped by the trimmed reads by using 'Map to Reference'

function with default settings except iterating up to 100 times in order to extend

17

sequences from the breaking points of the genome pieces. After generating consensus

from the mapping result of each fragments cut from the draft genome, these extended

pieces (consensus of the mapping) were mapped back to the draft genome, from which

the pieces were cut. The extensions from the cutting sites would form overlapping

sequences after the mapping, recovering the correct sequence of that species. Lastly,

the trimmed cp reads were mapped back to the corrected plastome sequence to check

overall correctness and to estimate the average coverage of each plastomes.

The IR/SC boundaries of each species were determined by mapping the cp reads

to the extracted IR sequence of Bhcp genome. The boundary of IR/SC become apparent

because of the different ends (5' and 3' ends) of the SC represented by the reads will

show two types of distinct sequence contents even if the read mapping of the query

species shows different boundary position from the reference IR (Fig. 4).
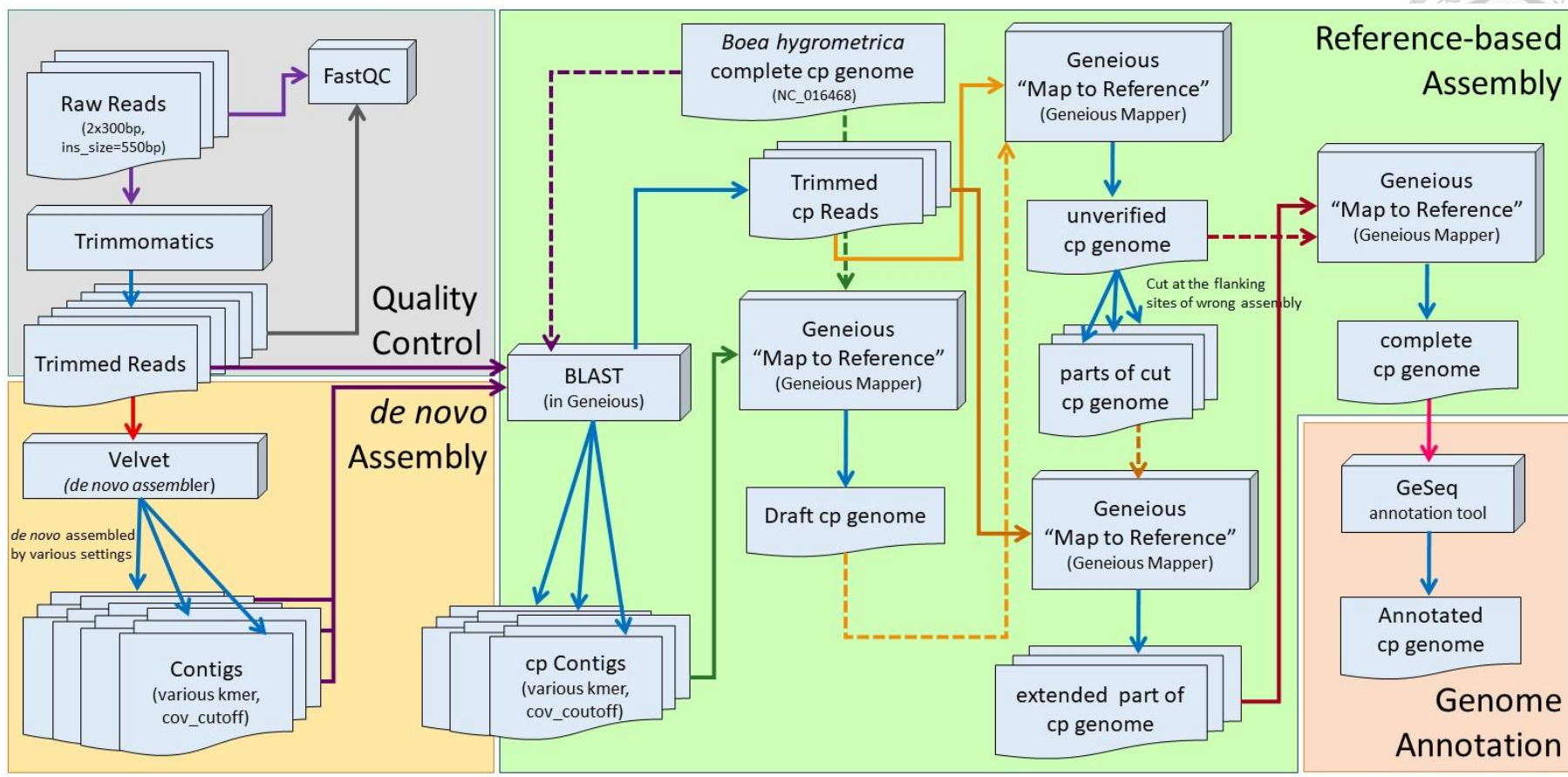
Fig. 2. **A brief workflow of plastome assembly.**

Blue arrow indicates the output of each steps. The rest of the different colors of arrows indicate each different step of assembly. The data served as reference sequence(s) for read mapping or blasting were connected with dash line to the tools that it were input. The boxes shown in 2D indicate the various files of sequences or contigs. The boxed shown in 3D indicate different analyzing tools.
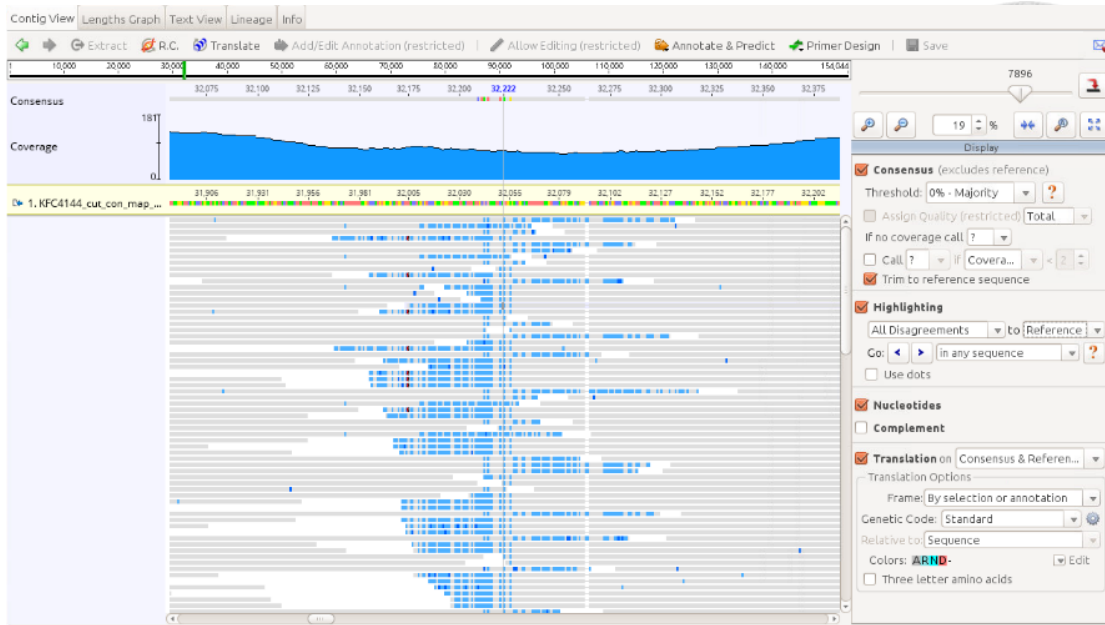
Fig. 3. The chimeric alignment of read mapping in Geneious mapper.

The chimeric alignment was caused by the large indels between the query species and the reference sequence. The disagreements between the reads and reference were highlighted by blue blocks.
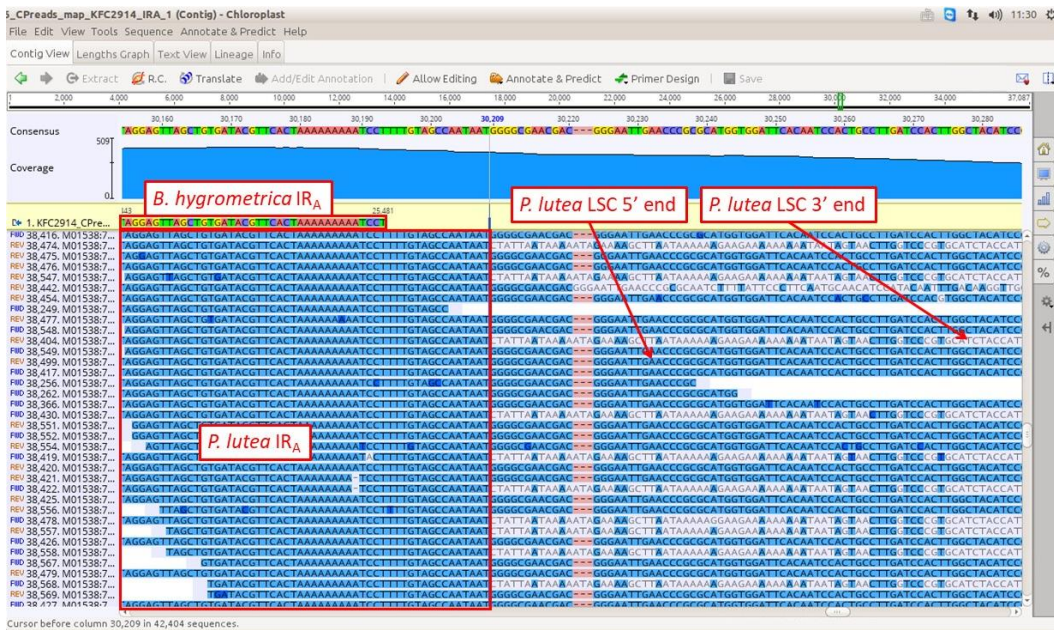
Fig. 4. Identifying the IR/SC boundary by mapping reads to the extracted reference sequence of IRA.

The IRA sequence was extracted and served as a reference to be mapped by the reads of the species whose IR/SC boundaries were not sure yet. By mapping the reads back to the extracted IR sequence, the boundary of the query species would be clear, for the different ends of SC will be aligned together showing two distinct types of sequence contents.

20

2.4 Plastome annotation

Instead of the widely used plastome annotation tool, DOGMA (Wyman *et al.*, 2004), I adopted GeSeq (Tillich *et al.*, 2017) web tool which provides more customized settings than DOGMA to perform primary annotation, then using 'transfer annotation' function in Geneious for the secondary annotation. The GeSeq annotator also integrates several third-party tRNA annotation tools which may be needed after the DOGMA annotation to confirm the presence of tRNA genes.

First, I chose two plastomes (*P. lutea* and *Pet. multiflorus*) representing two genera to be annotated by GeSeq with the following settings: BLAT protein search identity= 90; BLAT rRNA, tRNA, DNA search Identity= 95; 3rd party tRNA annotators, tRNAscan-SE v1.3.1 (Lowe and Eddy, 1997) activated with Mitochondria/Chloroplast Search Mode and default settings. Among the reference plastomes available in GeSeq server, nine were chosen containing five diverged species adopted in DOGMA and four Lamiales, to which is the order that Gesneriaceae belongs, species (Table 2).

Next, the results of annotation were visualized by viewing the genome maps output by GeSeq, which integrated OrganellarGenomeDRAW (OGDRAW; Lohse *et al.*, 2013) to construct the figure. The newly generated .gb files from GeSeq were downloaded and imported into Geneious to check gene boundaries and to delete the redundant annotations by mapping the extracted genes of Bhcp genome. To verify the correct

anticodon names and length of each tRNA genes, sequences of each tRNA genes were

searched in the online server version of tRNAscan-SE (Lowe and Chan, 2016).

Furthermore, genes with extremely short exons (*petB, petD*) or the notable trans-

splicing genes (*rps12*), which were often annotated erroneously (Tonti-Filippini *et al.*,

2017), were checked and annotated carefully by manual curation under Geneious. After

the plastomes of *P. lutea* and *Pet. multiflorus* being completely annotated, they were

mapped to each of the remaining unannotated plastomes of the same genera, and the

'transfer annotation' function in Geneious was applied to transfer annotation from the

annotated genome to the unannotated one. Finally, the annotated plastomes of every

species were aligned by MAFFT (Katoh *et al.*, 2002) with default settings in Geneious,

and the coding regions were shown in translated amino acid sequence to check the

correctness of annotations. By inputting the Genbank format (.gb) files generated by

Geneious, the genome maps were drawn by OGDRAW (Lohse *et al.*, 2013).

Table 2. Reference plastid genomes adopted in GeSeq annotating steps

Accession No.	Order	Family	Species	Reference	DOGMA*
NC_007407	Acorales	Acoraceae	*Acorus gramineus*	Goremykin *et al.* (2005)	+
NC_002693	Myrtales	Onagraceae	*Oenothera elata subsp. hookeri*	Hupfer *et al.* (2000)	+
NC_002202	Caryophyllales	Amaranthaceae	*Spinacia oleracea*	Schmitz-Linneweber *et al.* (2001)	+
NC_002694	Fabales	Fabaceae	*Lotus japonicus*	Kato *et al.* (2000)	+
NC_001879	Solanales	Solanaceae	*Nicotiana tabacum*	Shinozaki *et al.* (1986)	+
NC_016468	Lamiales	Gesneriaceae	*Boea hygrometricum*	Zhang *et al.* (2012)	
NC_013707	Lamiales	Oleaceae	*Olea europaea*		
NC_022451	Lamiales	Acanthaceae	*Andrographis paniculata*	Ding *et al.* (2016)	
NC_016433	Lamiales	Pedaliaceae	*Sesamum indicum*	Yi and Kim (2012)	

*The reference genome that was adopted in DOGMA was marked as '+'.

2.5 Plastome comparison

To compare the plastome sequences of different Gesneriaceae species, sequence files of 18 Gesneriad plastomes were uploaded to mVISTA (Frazer *et al.*, 2004) website to generate the similarity plot under the Shuffle-LAGAN mode which is able to improve the alignment of distant homologues. The sequence as well as the annotations of *P. lutea* were used as the reference for the similarity comparison. The plot was further edited in Adobe Illustrator CC (Adobe Systems, Inc., USA). The IRB regions of each plastomes in the alignment plot were removed to reduce redundancy.

The locations of the IR/SC boundaries and each genes flanking the boundaries were checked under the visualized interface of Geneious. The comparison of IR/SC boundaries of each Gesneriad plastomes were drawn in graph by using Adobe Illustrator CC (Adobe Systems, Inc.).
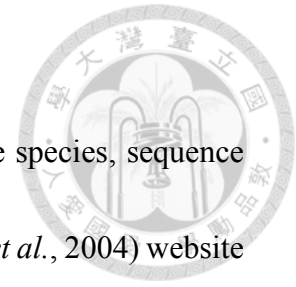
2.6 Plastome alignment, sequence divergence comparison and identification of variable regions
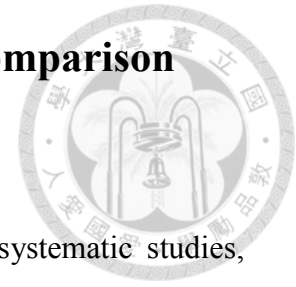
. To develop molecular markers applicable to intrageneric systematic studies, especially for resolving relationships among *Primulina* species, only the plastomes of *Primulina* and *Petrocodon* were included in this analysis. The ten successfully recovered plastomes of *Primulina* and *Petrocodon* generated in this study plus five *Primulina* plastomes downloaded from GenBank were aligned by MAFFT (Katoh *et al.*, 2002) with default settings in Geneious. The sequence divergence of 16 plastid genomes were estimated by the sliding window analysis performed in DnaSP 6 (Librado and Rozas, 2009) with window length of 800 bp and step size of 200 bp. The output data were plotted by Microsoft Excel with Pi value as y-axis. By referring to the method of Thomson *et al.* (2017), the regions with the windows of Pi values over 95th percentile (Pi value higher than or near to 0.02) were marked as variable regions, which are potentially useful for phylogenetic analyses.

2.7 Plastome repeat analyses
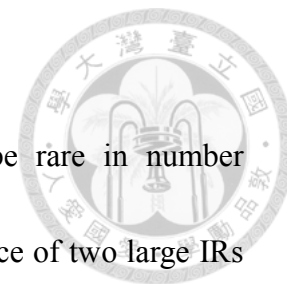
The repeat contents of plastome is often considered to be rare in number comparing to the nuclear genome with the exception of the presence of two large IRs (Palmer, 1991). Many of them have been proven to be useful in population genetics to evaluate the divergence between populations or close related species (e.g., Powell *et al.*, 1995; Cato and Richardson, 1996; Bryan *et al.*, 1999; Provan *et al.*, 2001). Hence, I searched the ten successfully assembled plastomes of this study for two main categories of repeats, the simple sequence repeats (SSRs, or microsatellites with motif repeating in mono-, di-, tri-, tetra-, penta-, and hexa- nucleotide units) and the dispersed repeats. To avoid redundancy, the following repeat analyses were performed on the plastomes containing only one IR. Currently, no universal criterion for the definitions of SSRs or dispersed repeats. I adopted the most frequently used criteria for each analyses after reviewing several similar studies (e.g., Martin *et al.*, 2013; Ni *et al.*, 2016; Jiang *et al.*, 2017) .

The SSR contents were detected by the Perl script MISA (Beier *et al.*, 2017) with the threshold of ten repeat units for mononucleotide SSRs, five repeat units for dinucleotide SSRs, four repeat units for trinucleotide SSRs, and three repeat units for tetra-, penta-, and hexanucleotide SSRs.

The dispersed repeats, which can be further divided into direct, reverse and palindromic repeats, were identified by REPuter (Kurtz *et al.*, 2001) with the minimum repeat size of 30 bp, 90 % or greater repeat identity by setting Hamming distance of 3. The repeats overlapping to its own copy or to other repeats were excluded.

2.8 Phylogenomic analyses based on plastomes

The ten complete plastomes assembled in this study plus eight other Gesneriad plastomes downloaded from GenBank (Table 3) were aligned by MAFFT (Katoh *et al.*, 2002) with default settings in Geneious and exported as NEXUS file. The alignment file was then imported into jModelTest 2.1.10 (Guindon and Gascuel, 2003; Darriba *et al.*, 2012) to test the nucleotide substitution models with default settings. By referring to the Akaike information criterion (AIC), the GTR+GAMMA+I model was suggested. The phylogenetic analysis based on maximum likelihood (ML) method using the model selected by jModelTest was performed by RAxML (Stamatakis, 2014) in Geneious with 'rapid bootstrapping and search for best-scoring ML tree' and bootstrap analysis with 500 replications to estimate branch support. The only European species, *Haberlea rhodopensis*, in the alignment was set as outgroup in the analysis. The phylogeny was drawn via FigTree v1.4.2.

27

2.9 Mitochondrial exome assembly, alignment and phylogenetic reconstruction
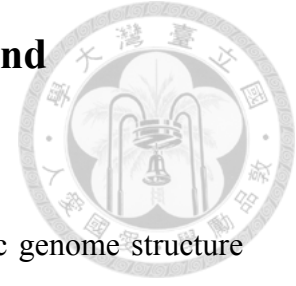
Plant mitochondrial genomes are well known for its dynamic genome structure and the larger genome size with low portion of coding regions comparing to plastid (Alverson *et al.*, 2010; Mower *et al.*, 2012). These characteristics have made the assembly of mitochondrial genome a difficult task. For genome skimming data, despite the recovery of complete mitochondrial genome being impossible, it is not difficult to retrieve the coding sequences alone (Straub *et al.*, 2012). The mitochondrial part still provides another unexplored information.

I tried to obtain mitochondria genes by mapping the filtered, non-plastid reads to each of the extracted exons separately from *B. hygrometrica* mitochondrial (Bhmt) genome (Zhang *et al.*, 2012), which has been the most close-related reference. After read mapping and checking start and stop codons of each genes, the consensus sequences of each exons were extracted. All the consensus sequences were concatenated in the order of the name of the exons and genes.

The concatenated exomes of each species, together with the concatenation of extracted CDS from Bhmt genome, were aligned by MAFFT with default settings in Geneious and exported to jModelTest to test the substitution model. The best model selected by the previous analysis were chosen for reconstructing ML phylogenetic tree

28

by using RAxML in Geneious with the settings as those performed in plastome

phylogenetic analyses, but the alignment of mt exomes being further partitioned by

three different codon positions. The concatenated Bhmt exome was set as outgroup of

the analysis. The output tree file was drawn into graph by FigTree v1.4.2.
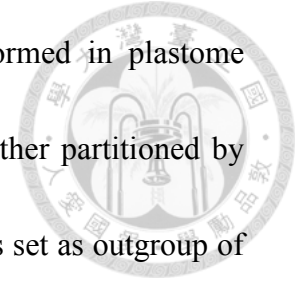
2.10 Nuclear ribosomal DNA assembly, alignment and phylogenetic analyses
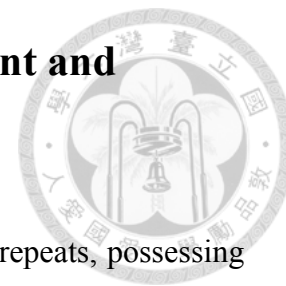
Nuclear ribosomal DNA (nrDNA) usually appears in tandem repeats, possessing multiple copies in plant genomes (Appels and Honeycutt, 1986; Rogers and Bendich, 1987) which makes it easy to be recovered under the genome skimming method. The two internal transcribed spacers (ITS1 and ITS2) between three ribosomal subunits genes of nrDNA have been another frequently used markers for plant systematics other than plastid DNA (Álvarez and Wendel, 2003). I tried to recover the complete cistron of nrDNA, which contains 18S, 5.8S, 26S genes, and ITS1, ITS2 sequences, aiming to reconstruct a phylogeny representing the nuclear genealogy.

The ITS sequence of *P. pengii* (GenBank accession number: KF498139) were chosen as the reference sequence to be mapped by the cp-excluded reads of *P. pengii* by Geneious mapper ('map to reference' function) with the same setting as the assembly of mt exomes. Since the18S and 26S gene of the ITS sequence of *P. pengii* were not completely sequenced, the ribosomal genes in the result of read mapping cannot be fully annotated. The nrDNA sequence with full length of ribosomal genes of *Perilla citriodora* (Nakai) Makino (GenBank accession number: KT220699) was mapped to the consensus extracted from the read mapping results of *P. pengii* to annotate the ribosomal genes. After the annotation, the complete nrDNA cistron of *P. pengii* was

30

extracted and then served as a reference for the rest of the samples for nrDNA cistron assembly.
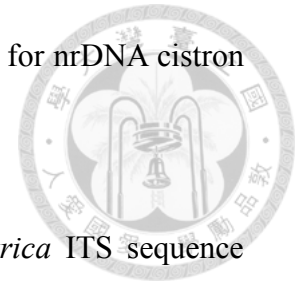
The DNA of studied species together with the *B. hygrometrica* ITS sequence (GenBank accession number: KU203783) were aligned by MAFFT in Geneious with default setting. The alignment was imported to jModelTest to test the appropriate nucleotide substitution models by inferring to AIC. The ML phylogenetic tree was reconstructed by RAxML under Geneious with rapid bootstrapping of 500 times and *B. hygrometrica* was set as an outgroup. The resulting tree file was further visualized by FigTree.

31
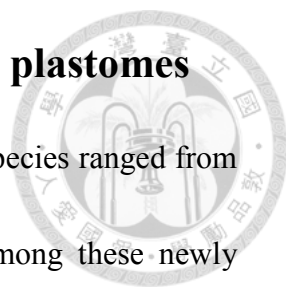
3. Results

3.1 Sequencing and assembly of plastomes

Among the 11 species being sent to library preparation and sequencing, the output of *P. lungzhouensis* was almost empty. This might result from the low quantity and poor quality of this DNA sample. Except the failed one, the remaining ten samples have been sequenced successfully with output data size ranging from 0.7 – 2.1 Gb. By calculating the number of the reads mapped to assemble the plastome, the average coverage of each plastome assembled were estimated. The coverage of most of the assembled plastomes were over 100× with two species being about 70× (Table 1).
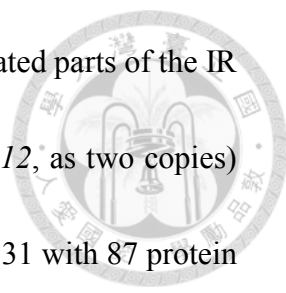
The proportion of plastid DNA to the total extracted genomic DNA were also estimated by recording the number of cp reads mapped to the complete plastomes of the same species divided by the number of the total trimmed reads. The cp reads of *Primulina* and *Petrocodon* species accounted for c.a. 5% of the genomic DNA (with the range of 0.96 – 6.3%) (Table 1). It is basically congruent with the results from several studies about plastid DNA proportion in a cell (Bendich, 1987; Straub *et al.*, 2012). However, the amount of cpDNA varies considerably among different stages of leave development and the cell types (Rauwolf *et al.*, 2009; Rowan *et al.*, 2009; Straub *et al.*, 2012). The proportion of cpDNA to nuclear DNA is an important consideration for increasing the multiplexing level of genome skimming data (Straub *et al.*, 2012).

32

3.2 Characteristics of *Primulina* and *Petrocodon* plastomes

Sizes of the complete plastomes of the ten newly sequenced species ranged from 152323 bp (*P. medica*) to 153786 bp (*P. cordata*) (Table 3). Among these newly generated plastomes, the length of LSC ranging from 83218 bp (*P. pengii*) to 84890 bp (*P. medica*), the length of IRs ranging from 25415 bp (*Pet. multiflorus*) to 25495 bp (*P. pengii*), and the length of SSC ranging from 16489 bp (*P. medica*) to 18279 bp (*Pet. multiflorus*) (Table 3). The GC contents of assembled plastomes ranged from 37.5 − 37.6 % indicating the obviously AT-biased sequence content of the plastome which is consistent with most of the plant species (Shimada and Sugiura, 1991). However, the GC contents of IR are significantly higher than the values in SC regions (Fig. 5 and Table 3).

In the plastomes of *Primulina* and *Petrocodon,* no pseudogene, which is defined as genes possessing immature stop codon(s), was detected. The gene contents of the eight *Primulina* and two *Petrocodon* plastomes are identical. They harbor in total of 111 different genes, including 80 different protein coding genes, 27 different tRNA genes, and four different rRNA genes (Table 4). Among the 111 genes, 19 were completely duplicated in IR and three (*rps19*, *ndhF* and *ycf1*) were partially duplicated resulting from their locations on IR/SC boundaries. The trans-splicing genes, *rps12*, were partially duplicated with two of its exons located in IRs. Including those genes

33

completely duplicated in IRs and excluding the incompletely duplicated parts of the IR

boundaries located genes (but counting the trans-splicing gene, *rps12*, as two copies)

the overall gene number of a *Primulina* or *Petrocodon* plastome is 131 with 87 protein

coding genes, 36 tRNA genes and 8 rRNA genes (Table 4). There are 16 intron

containing genes (not counting duplicated copy) in total with 11 protein coding genes

and five tRNA genes (Table 4). Only the *ycf3* and *clpP* gene contain two introns while

the remaining genes contain one intron (Table 4). There are three pairs of genes

overlapping with each other (*psbD*/*psbC*, *ndhK*/*ndhC*, *atpE*/*aptB*). The length of the

coding regions (including RNA genes) accounts for 44.2 % of the total size of a

Primulina or *Petrocodon* plastome in average, indicating the plastome structure of

species sequenced in this study is relatively compact comparing to the nucleus whose

genome is usually dominated by non-coding sequences (Schmidt and Heslop-Harrison,

1998; Vicient *et al.*, 1999; Meyers *et al.*, 2001; Feschotte *et al.*, 2002).

By comparing the gene contents of the two genera with the reference Bhcp genome

and other Gesneriad plastomes, regardless of the annotation disparity which means the

same gene was annotated into different names, there is no gene losses, relocations or

structural inversions. This comparison suggest that plastomes of Gesneriaceae species

sampled here are highly conserved in gene contents and gene orders.

34
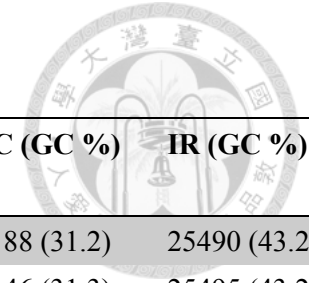
Table 3. Plastome features of Gesneriaceae species.

Scientific name	Chinese name	Voucher No./ *Accession No.*	cp size (GC %)	LSC (GC %)	SSC (GC %)	IR (GC %)
Primulina lutea	黃花牛耳朵	KFC2916	153262 (37.6)	84094 (35.6)	18188 (31.2)	25490 (43.2)
Primulina pengii	彭氏報春苣苔	Peng24024	152354 (37.7)	83218 (35.7)	18146 (31.3)	25495 (43.2)
Primulina eburnea	牛耳朵	*MF177038*	152963 (37.6)	83934 (35.6)	18099 (31.1)	25465 (43.2)
Primulina longgangensis	弄崗唇柱苣苔	Peng22948	153191 (37.7)	84105 (35.6)	18168 (31.3)	25459 (43.2)
Primulina linearifolia	線葉唇柱苣苔	*MF472013*	153244 (37.6)	84149 (35.6)	18177 (31.2)	25459 (43.2)
Primulina liboensis	荔坡唇柱苣苔	*MF177039*	152989 (37.7)	84585 (35.6)	17478 (31.4)	25463 (43.2)
Primulina sclerophylla	硬葉唇柱苣苔	KFC2971	153185 (37.6)	84608 (35.6)	17641 (31.3)	25468 (43.2)
Primulina hezhouensis	賀州小花苣苔	KFC2914	153556 (37.6)	84757 (35.5)	17887 (31.3)	25456 (43.2)
Primulina huaijiensis	懷集報春苣苔	*MF472012*	153401 (37.6)	84334 (35.6)	18213 (31.2)	25427 (43.2)
Primulina medica	藥用唇柱苣苔	KFC2962	152323 (37.6)	84890 (35.5)	16489 (31.4)	25472 (43.2)
Primulina fimbrisepala	螞蝗七	KFC4144	153089 (37.5)	84684 (35.4)	17527 (31.2)	25439 (43.2)
Primulina cordata	心葉唇柱苣苔	KFC3021	153786 (37.5)	84789 (35.5)	18085 (31.2)	25456 (43.2)
Primulina brachytricha var. *magnibracteata*	短毛唇柱苣苔	*MF177037*	153723 (37.5)	84783 (35.4)	18078 (31.2)	25431 (43.2)
Petrocodon coriaceifolius	革葉石山苣苔	KFC2943	153292 (37.5)	84345 (35.4)	18083 (31.2)	25432 (43.2)
Petrocodon multiflorus	多花石山苣苔	KFC2913	153227 (37.5)	84118 (35.4)	18279 (31.0)	25415 (43.2)
Lysionotus pauciflorus	吊石苣苔(石吊蘭)	*KX752081*	153856 (37.5)	85087 (35.4)	17839 (31.2)	25465 (43.2)
Boea hygrometrica	旋蒴苣苔	*NC016468*	153493 (37.6)	84701 (35.6)	17900 (31.0)	25446 (43.2)
Haberlea rhodopensis		*KX657870*	153254 (37.7)	84299 (35.7)	17837 (31.7)	25559 (43.2)

Note: The unit of size and length of each plastomes is base pair (bp). The accession numbers of sequences from GenBank were shown in italic.
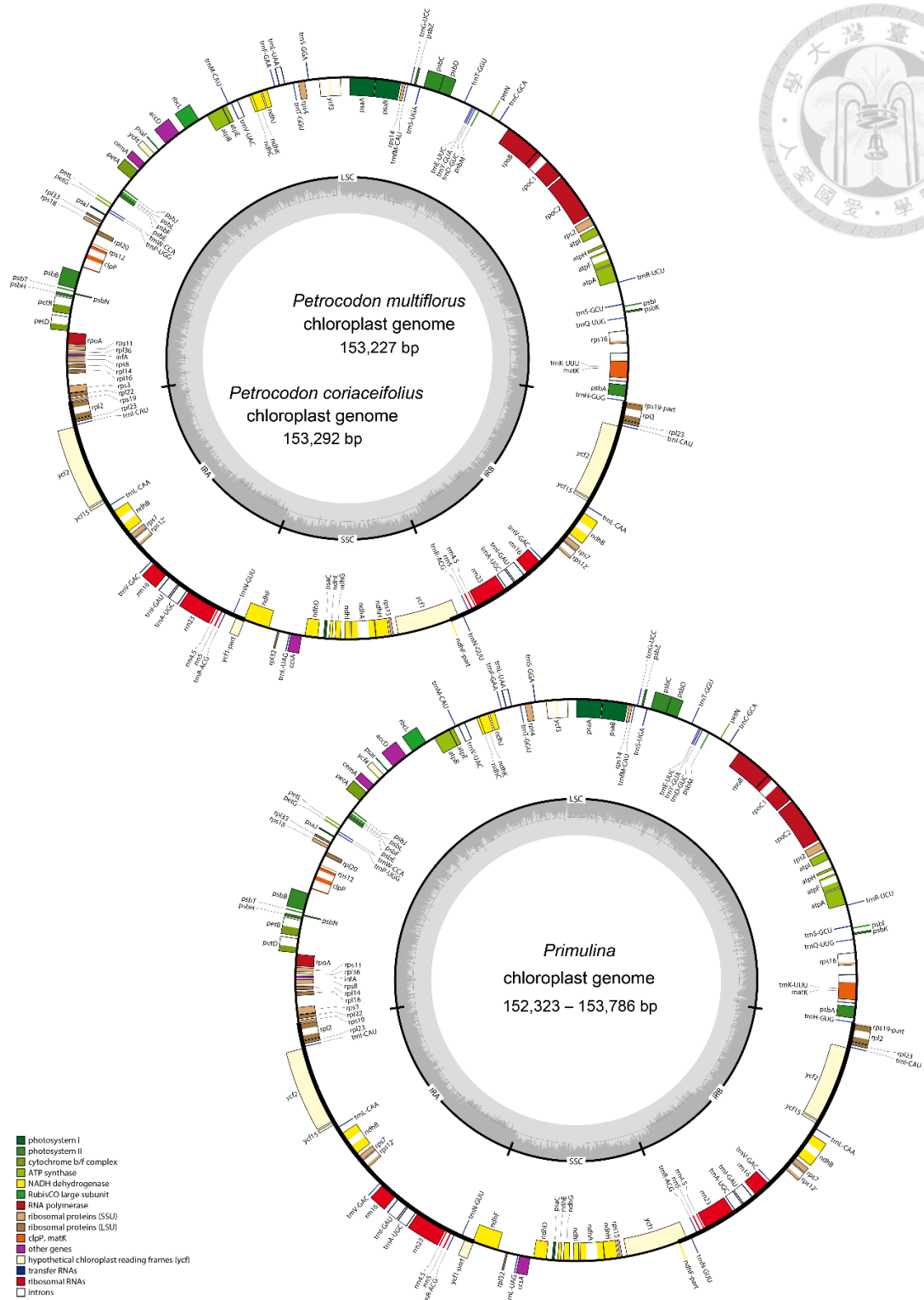
Fig. 5. The genome maps of *Primulina* and *Petrocodon* plastomes.

The genes drawn on inner side of the outer circle are transcribed clockwise, and those on the outer side are transcribed counterclockwise. IRs are shown in bold line in the outer circle. The inner circle indicates GC contents across the genome with lighter gray indicating AT contents. Genes belonging to different functional groups were shown in different colors.

36

Table 4. The gene contents of *Primulina* and *Petrocodon* plastomes.

Gene category	Gene Name
Photosystem I	*psaA, psaB, psaC, psaI, psaJ*
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	*atpA, atpB, atpE, atpF*, atpH, atpI*
NADH dehydrogenase	*ndhA*, ndhB(×2)*, ndhC, ndhD, ndhE, ndhF#, ndhG, ndhH, ndhI, ndhJ, ndhK*
RubisCO large subunit	*rbcL*
RNA polymerase	*rpoA, rpoB, rpoC1*, rpoC2*
Ribosomal proteins (SSU)	*rps2, rps3, rps4, rps7(×2), rps8, rps11, rps12(×1.67)†*, rps14, rps15, rps16*, rps18, rps19#*
Ribosomal proteins (LSU)	*rpl2(×2)*, rpl14, rpl16, rpl20, rpl22, rpl23(×2), rpl32, rpl33, rpl36*
Protease, Maturase	*clpP**, matK*
Other genes	*accD, ccs1, ccsA, infA, lhbA, cemA*
Hypothetical chloroplast reading frames (ycf)	*ycf1#, ycf2(×2), ycf3**, ycf4, ycf15(×2)*
Transfer RNAs (tRNA)	*trnA-UGC(×2)*, trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnG-UCC, trnH-GUG, trnI-CAU(×2), trnI-GAU(×2)*, trnK-UUU*, trnL-CAA(×2), trnL-UAA*, trnL-UAG, trnM-CAU, trnfM-CAU, trnN-GUU(×2), trnP-UGG, trnQ-UUG, trnR-ACG(×2), trnR-UCU, trnS-GCU(×2), trnS-UGA, trnT-GGU(×2), trnV-GAC(×2), trnV-UAC*, trnW-CCA, trnY-GUA*
Ribosomal RNAs (rRNA)	*rrn4.5(×2), rrn5(×2), rrn16(×2), rrn2(×2)*

(×2): Genes duplicated in IRs.

*Genes with one intron.

**Genes with two introns.

† The trans-splicing gene, *rps12*, which possesses three exons with two of which duplicated in IRs.

Therefore, the gene was marked as ×1.67 copies.

The partially duplicated genes locating on the IR/SC boundary.

3.3 Plastome comparison and identification of variable regions

The alignment identity plot generated by mVISTA was shown in Fig. 6. The result indicates that the cp sequences are highly conserved among the 18 Gesneriaceae species included in this study. The sequences of IRs are even more conserved than the SC regions. This is consistent with the general knowledge about IR whose substitution rate is lower than LSC and SSC, because IRs are identical to each other, being able to repair mutations from other copies via biased gene conversion (Birky and Walsh, 1992), and to stabilize the overall plastid genome structure (Blazier *et al.*, 2016).

Since the indels (the alignment gaps) were not excluded in the mVISTA plot, some part of the graph seemed to be very variable consequently (Fig. 6).The variable regions in the mVISTA analyses may not be phylogenetically informative. I performed the sliding window analysis (Fig. 7) which evaluated the sequence divergence by Pi value using DnaSP. In total of eight regions, *trnS-trnR, rpoB-trnC-petN, psbZ-trnG, rps4-trnT-trnL, psaJ-rpl33, petD* intron, *rpl32-trnL*, and *ycf1*, whose Pi values were over 95th percentile, were identified as variable regions (Fig. 7).

38

Fig. 6. The mVISTA genome similarity plot of 18 Gesneriaceae plastomes.

P. lutea was used as a reference for the alignment and similarity comparison. The gaps and differences comparing to the reference were shown in blank in the plot. Sequence similarity under 50% was not shown here. Different colors indicate different sequence type. The alignment gaps and dissimilarities were shown in blank area in the plot.

39
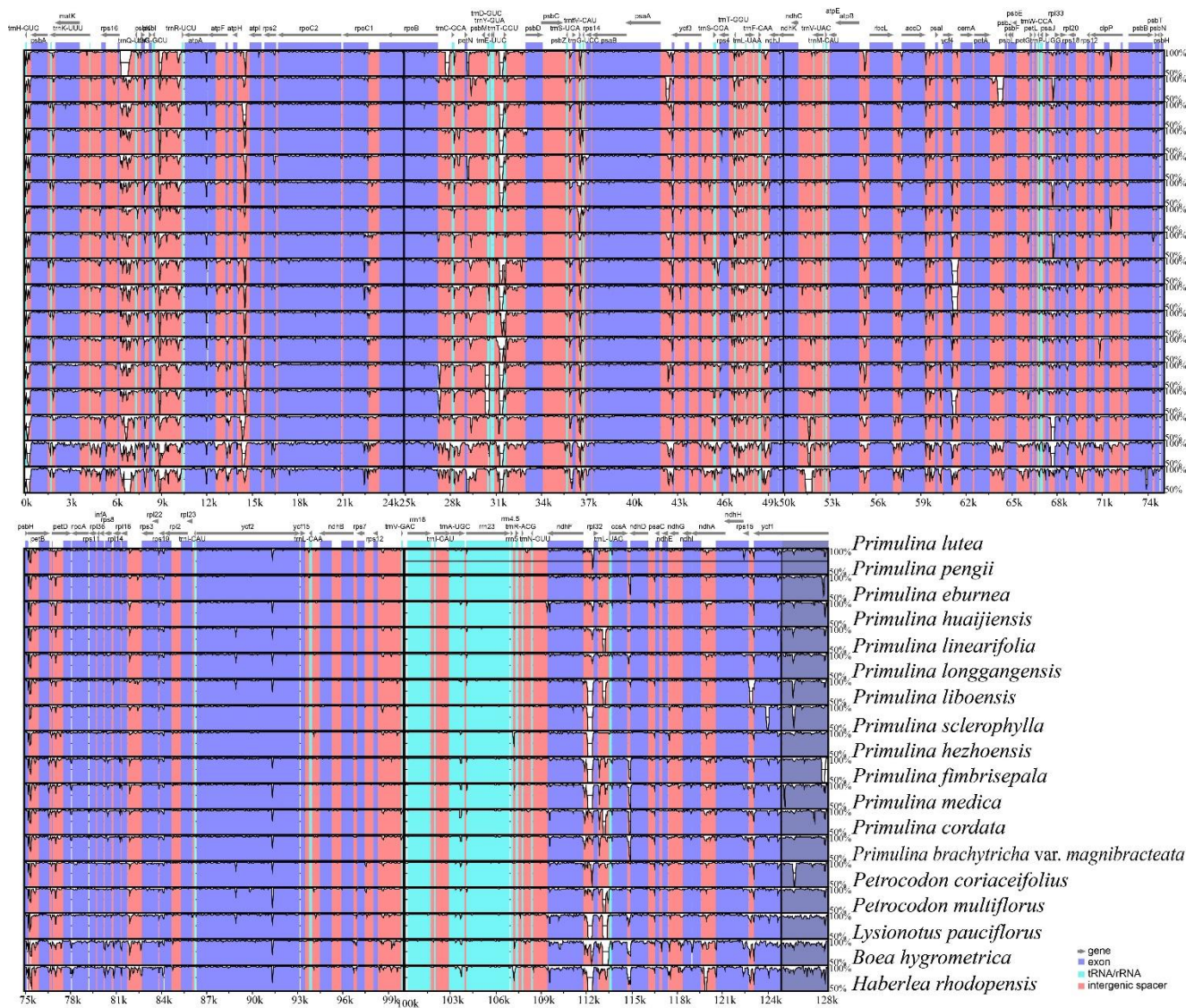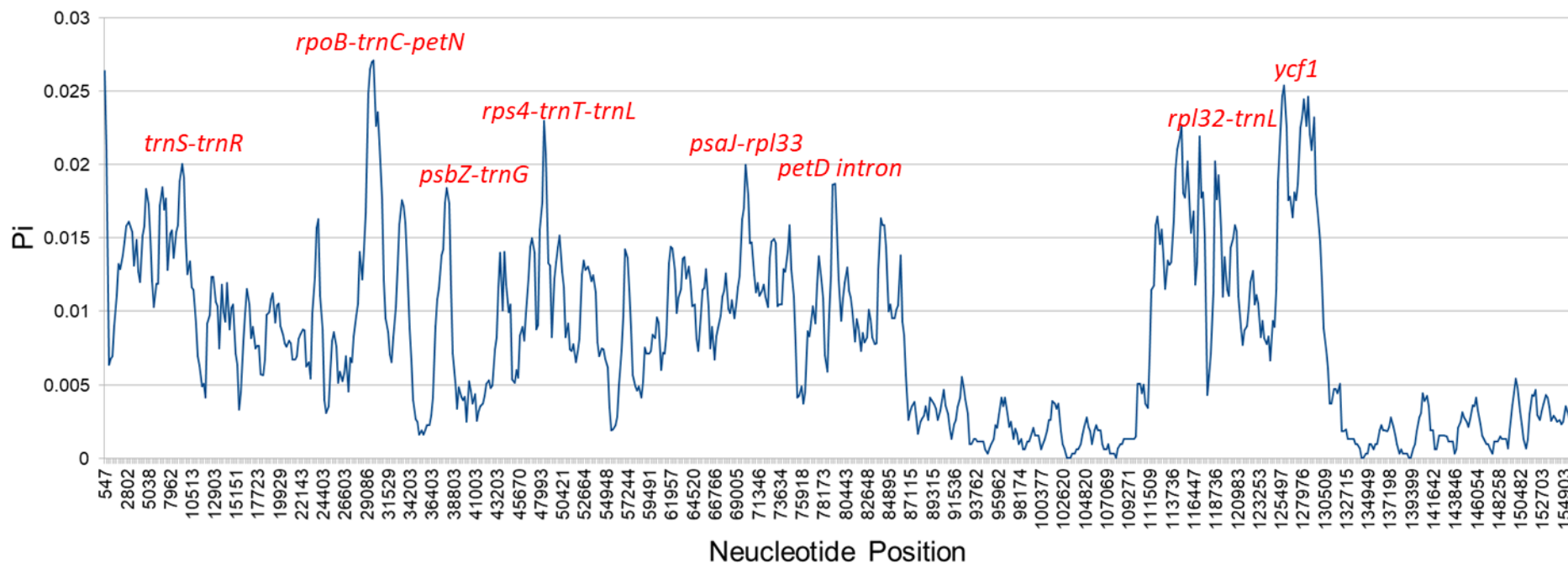
Fig. 7. The sliding window analysis of 15 *Primulina* and 2 *Petrocodon* plastomes alignment with variable regions designated.

3.4 IR/SC boundaries of Gesneriad plastomes
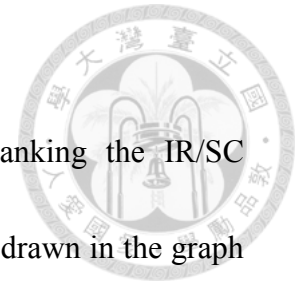
The IR/SC boundaries and the positions of the genes flanking the IR/SC boundaries of the Gesneriad plastomes included in this study were drawn in the graph (Fig. 8). The order of the studied species included in Fig. 8 were determined by their phylogenetic relationships based on the phylogenies inferred from complete plastomes (Fig. 11). The IR/SC boundaries among studied species are highly conserved with very minor expansions or contractions except for *Haberlea rhodopensis*, which is one of basal lineages in Gesnericaceae (Möller *et al.*, 2011), with its IRs contracting to exclude the *ndhF* gene despite its longer IRs. My data show that the more close-related the species are, the more minor the IR/SC boundaries shift. Moreover, the LSC/IR boundary is usually more conserved than the IR/SSC one, which is consistent in many plant taxa (Zhu *et al.*, 2016).
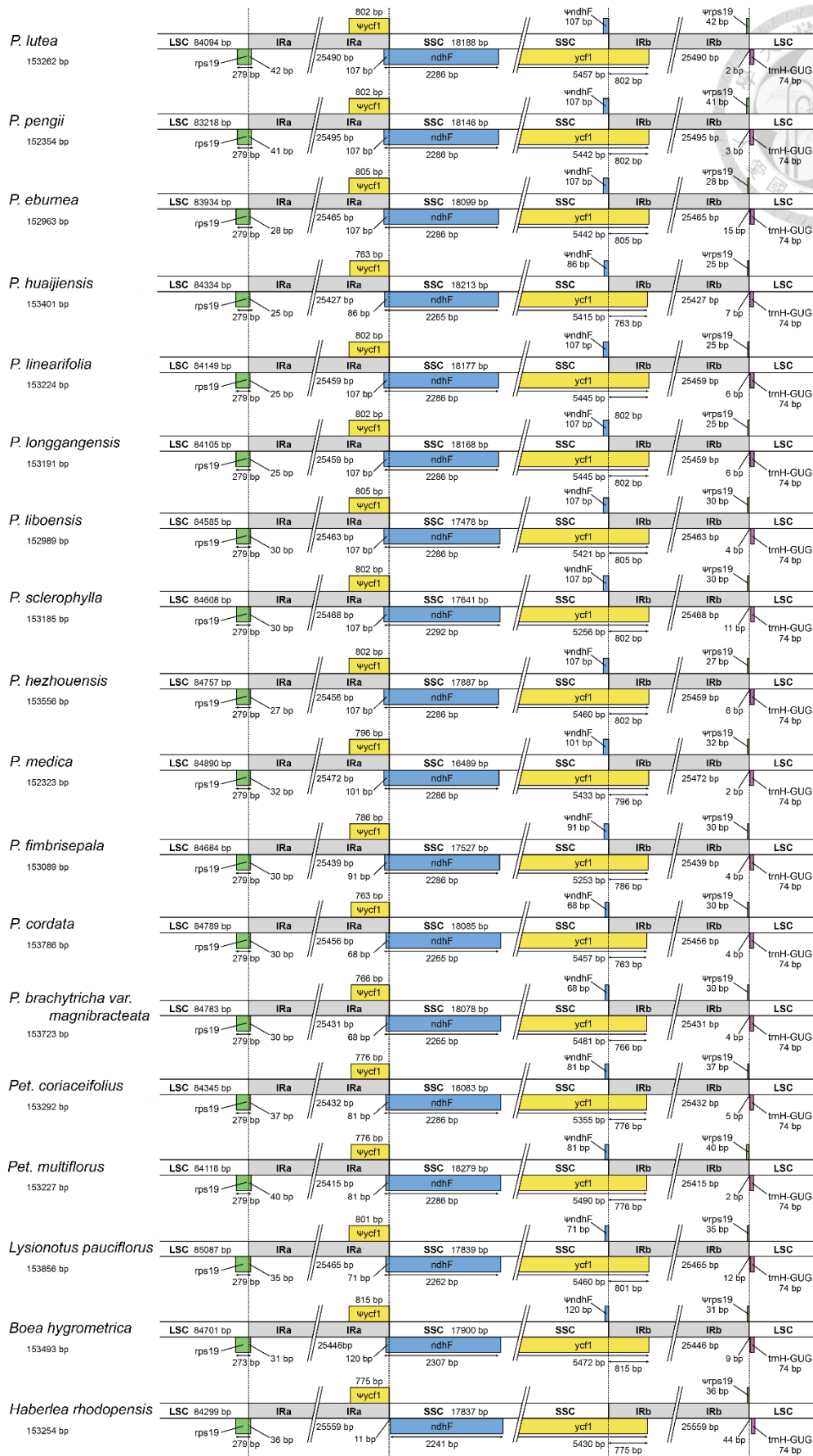
Fig. 8. The IR/SC boundary graph of 18 Gesneriaceae plastomes.

3.5 Repeat contents in *Primulina* and *Petrocodon* plastome

Despite the generally low repeat contents in plastid genomes (Palmer, 1991), a moderate amount of SSRs and few dispersed repeats were identified in *Primulina* and *Petrocodon* plastomes. The SSRs in chloroplast have been employed as useful genetic markers in population genetic studies owing to its high mutation rate and the features of uniparental inheritance and haploid genome (Powell *et al.*, 1995).

There are only 22 different dispersed repeats identified from the ten plastomes with over half of the repeats were presented in multiple or all of the plastomes (Table 5). The shared sequence contents of the disperse repeats have shown that the repeats in *Primulina* and *Petrocodon* are quite conserved in both sequence contents and locations of repeats in the genomes. Most of the dispersed repeats were in the length of 30 - 45 bp, and none of them was longer than 60 bp. Palindromic repeat is the most abundant type among the three different repeat classes (forward, reverse, palindromic; Table 5).

The number of SSRs in each plastomes ranged from 28 to 44 (Fig. 9). Most of the SSRs are mononucleotide A/T repeats with very few C/G types (Fig. 8) indicating that the cp SSRs are also strongly AT-biased. Over half of the SSRs are located in intergenic spacers (Fi.g 10 A). Most of the SSRs distribute in LSC regarding to the long length of LSC (Fig. 10 B). By contrast, IRs contain very few SSRs comparing to SSC which is shorter than IRs but sometimes contains more SSRs than IRs in my results (Fig. 10 B).

43

Table 5. The locations and lengths of dispersed repeats of the ten plastomes.

No. of copy	\[LSC\] rps16-trnQ	psbK-psbI	**trnS**	trnS-trnR	atpF-atpH	psbZ-trnG	**psaB**	psaA-ycf3	*ycf3 intron 2*	*ycf3 intron 1*	rps4-trnT	psaI-ycf4	*petB intron*	petA-psbJ	*clpP intron 2*	rpl16-rps3	**ycf2**	\[IR\] ycf15-trnL	rrn4.5-rrn5	\[SSC\] ndhF-rpl32
2nd			**trnS**				**psaA**			rps12'-trnV					ndhF-rpl32					
3rd										*ndhA intron*										
Length	32	31	30	35	30	30	44	32	87	41	31	52	41	33	31	31	30	35	30	31
Pmlt			+		+		+			+		+	+					+	+	
Pmlg			+		+	+	+	+(42)		+		+		+				+	+	
Pmpg	+		+		+	+	+			+		+		+				+	+	
Pmhz			+		+	+	+			+		+		+				+		
Pmsc			+		+		+			+		+		+				+	+	
Pmmd		+	+	+	+		+	+	+	+		+		+		+		+	+	
Pmfb			+		+		+			+		+		+			+(31)	+	+	
Pmcd			+		+	+	+			+	+	+		+				+	+	
Ptmt			+		+	+	+			+				+	+	+		+	+	+
Ptcc			+		+		+			+				+		+	+	+	+	+

Note: The locations of repeats shown in blue, red and black colors indicate forward, reverse and palindromic repeats, respectively. The repeat locations belonging to intergenic spacers, coding regions and introns were shown in regular, bold and italic letters, respectively. The presence of the repeat in the specific region was marked by '+' sign. Repeats in the same region without number marked in brackets indicates the identical repeats shared by different species. The repeats with different sequence content from the same regions of other species were further marked by the number which indicates the lengths (bp) of the unique repeat. Abbreviations: Pmlt: *Primulina lutea*; Pmpg: *P. pengii*, Pmlg: *P. longgangensis*; Pmhz: *P. hezhouensis*; Pmsc: *P. sclerophylla*; Pmmd: *P. medica*; Pmfb: *P. fimbrisepala*; Pmcd: *P. cordata*; Ptmt: *Petrocodon multiflorus*; Ptcc: *Pe. coriaceifolius*.
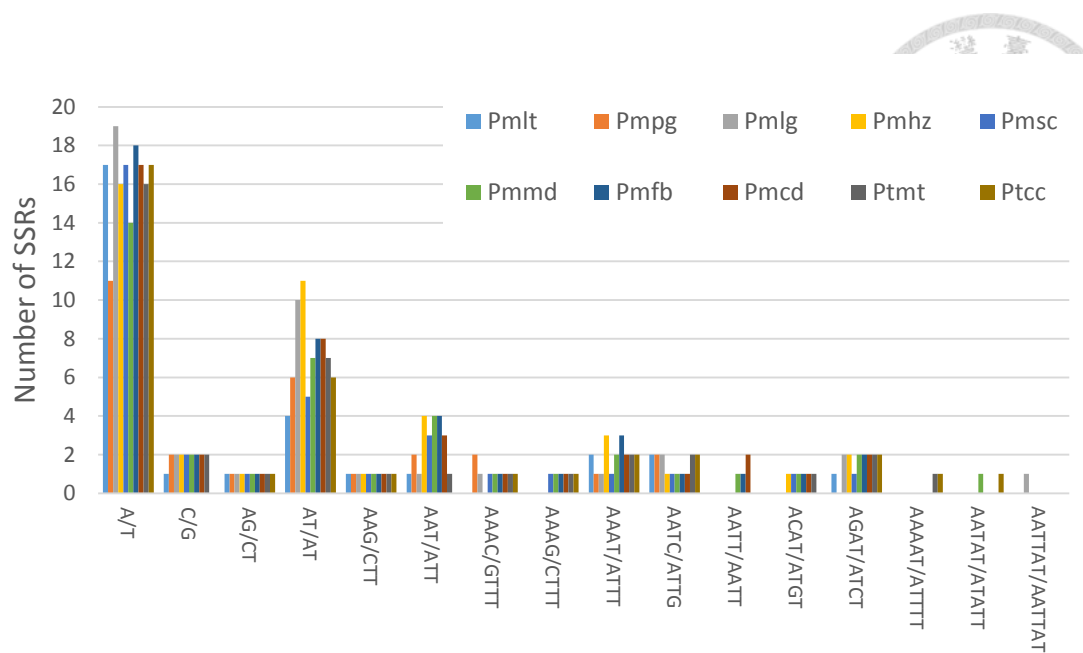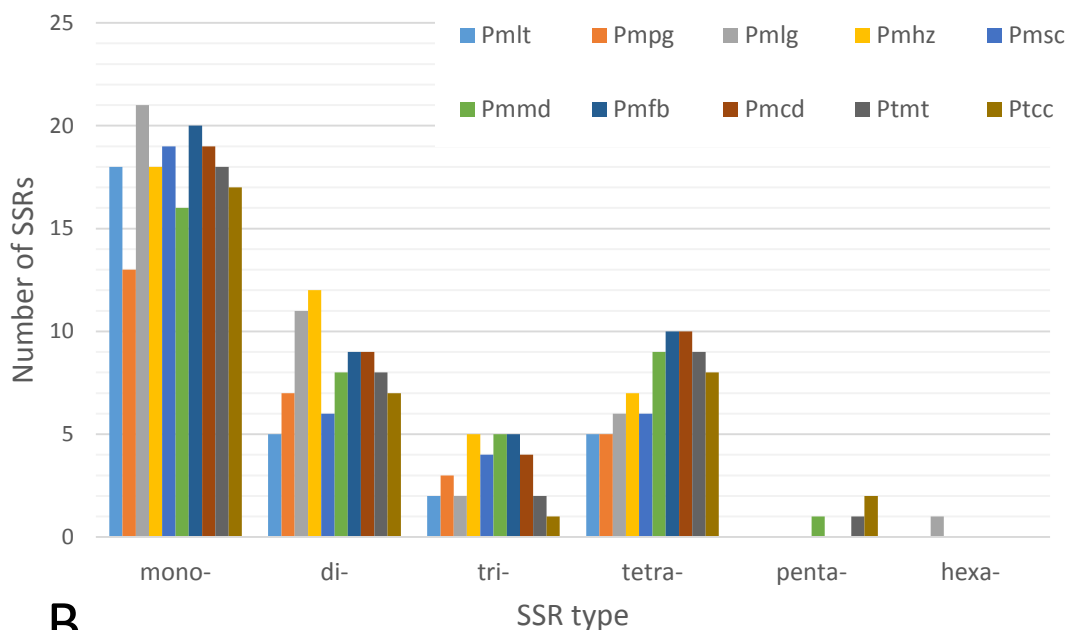
Fig. 9. Number of SSRs and its motif types in the ten newly sequenced plastomes.
A: The numbers of different motif sequence types of each plastomes. **B:** The numbers of different SSR type of each species.

Abbreviations: Pmlt: *Primulina lutea*; Pmpg: *P. pengii*, Pmlg: *P. longgangensis*; Pmhz: *P. hezhouensis*; Pmsc: *P. sclerophylla*; Pmmd: *P. medica*; Pmfb: *P. fimbrisepala*; Pmcd: *P. cordata*; Ptmt: *Petrocodon multiflorus*; Ptcc: *Pe. coriaceifolius*.

Fig. 10. **The numbers and locations of SSR of each species.**

A: SSR distribution in intergenic spacers (IGS), introns and coding sequences (CDS).

B: SSR distribution in three different part of the plastome.

Abbreviations: Pmlt: *Primulina lutea*; Pmpg: *P. pengii*, Pmlg: *P. longgangensis*; Pmhz: *P. hezhouensis*; Pmsc: *P. sclerophylla*; Pmmd: *P. medica*; Pmfb: *P. fimbrisepala*; Pmcd: *P. cordata*; Ptmt: *Petrocodon multiflorus*; Ptcc: *Pe. coriaceifolius*.
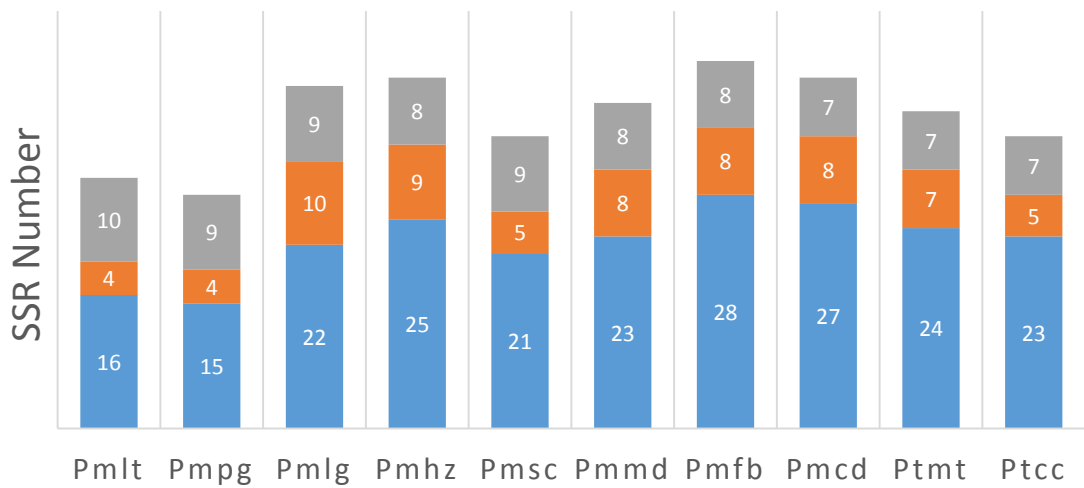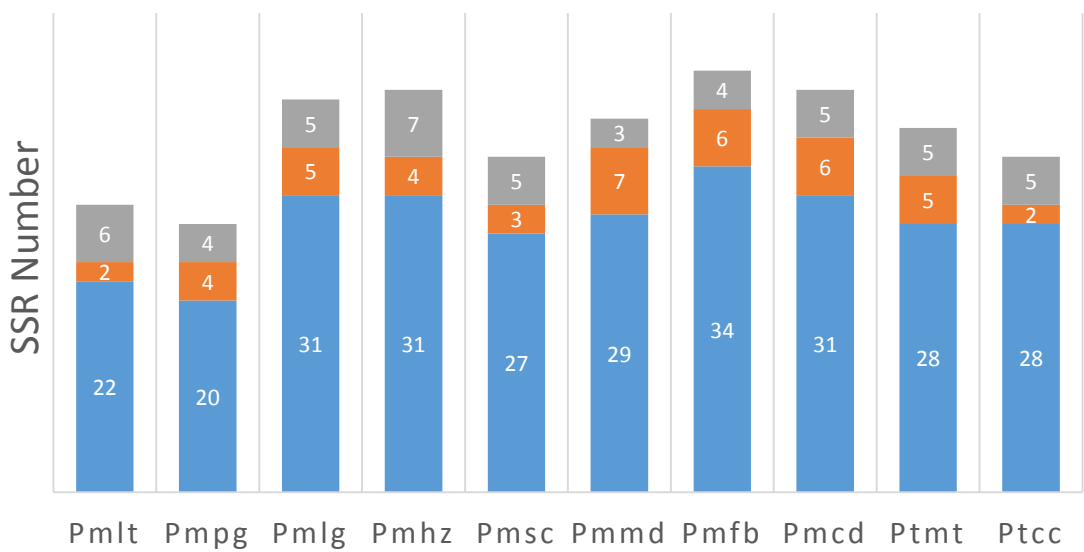
3.6 Chloroplast phylogenomic analyses

The phylogenetic relationships among studied species were resolved in extremely high support value with most of the nodes are 100% bootstrap supported (Fig. 11). The tree topology of this phylogeny is largely congruent with the results from Guo *et al.* (2015) with few minor differences. The designations of the major *Primulina* clades in my results (Fig. 11) were referred to Guo *et al.* (2015) with one additional clade C (*P. sclerophylla, P. liboensis, P. hezhouensis*; Fig. 11) assigned by my study.
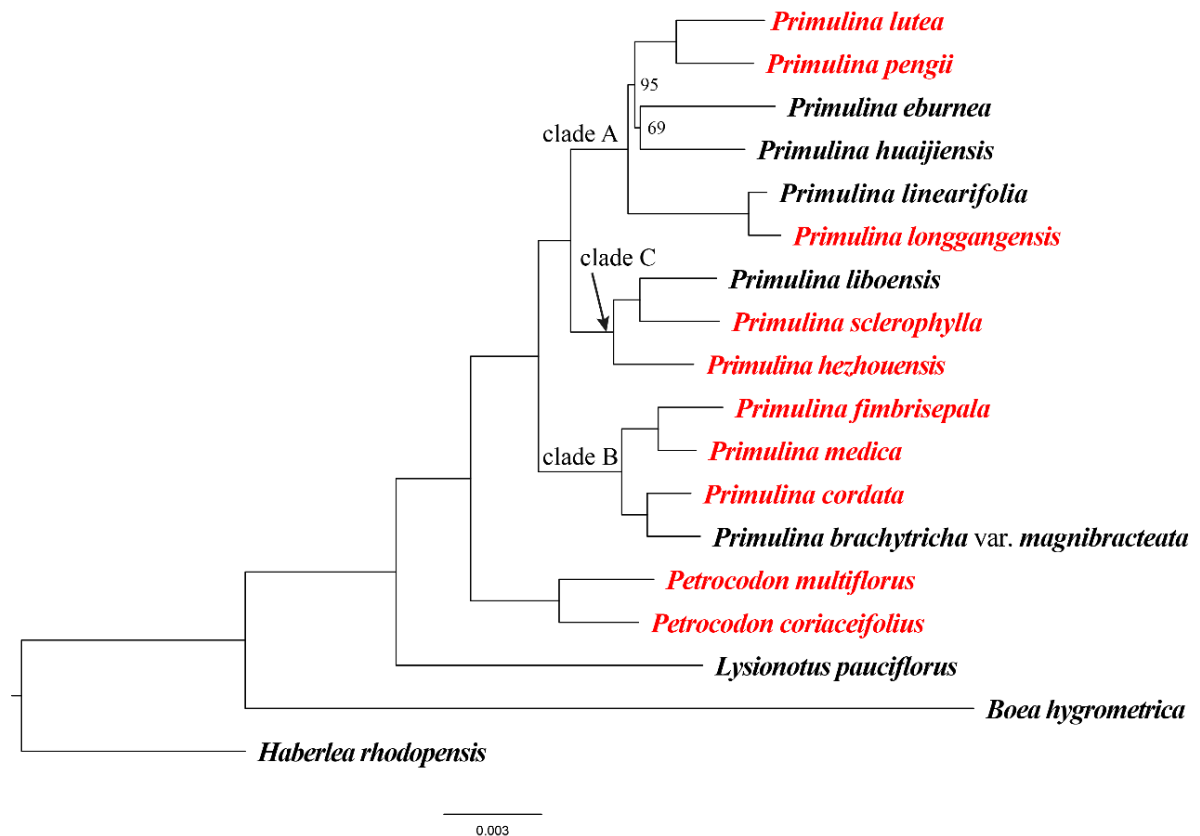
Fig. 11. The ML phylogeny inferred from complete plastomes of 18 Gesneriaceae species.
Only the bootstrap values <100% were shown in the graph. The remaining nodes are with 100% bootstrap support. The species newly sequenced were marked in red color. The designation of clade A and clade B was referred to Guo *et al.* (2015).

3.7 Assembly and phylogenetic analysis of mitochondrial exome
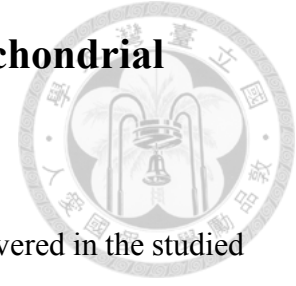
Not all the genes extracted from the Bhmt genome can be recovered in the studied *Primulina* and *Petrocodon* species. Among the 29 protein coding genes (51 exons in total) in Bhmt genome, *atp1*, *apt6*, *nad5* exon 3 and *rps7* genes were failed in assembly because of erroneous assemblies or empty reads found to be mapped to the reference sequences. Therefore, these regions were excluded from further analyses. Exons which were fully recovered without incorrect assembly and were adopted in the subsequent analyses are listed in Table S1.

The alignment length of the concatenated mitochondrial coding sequences is 24957 bp. Despite the large alignment, there were only 1.8% of variable sites, and the phylogenetic relationship based on the data shows lower bootstrap values than the phylogeny from cp data (Fig. 12).

Fig. 12. The ML phylogeny inferred from 26 mitochondrial protein coding genes of 11 Gesneriaceae species.
Each node was labelled with bootstrap support.

48

3.8 Phylogenetic analysis of nuclear ribosomal sequences

The phylogeny inferred from ITS and ribosomal sequences shows moderate to strong support of each clades (Fig. 13). The length of nrDNA alignment is 5939 bp with 87.6% of identical sites. The overall resolution and support value is lower than the cp and mt phylogenies because of the shorter alignment and fewer informative sites than the cp and mt alignments (Fig. 12, 13). *Primulina longgangensis, P. sclerophylla* and the clade containing *P. lutea* and *P. pengii* form a polytomy with very low statistical support (Fig. 13).
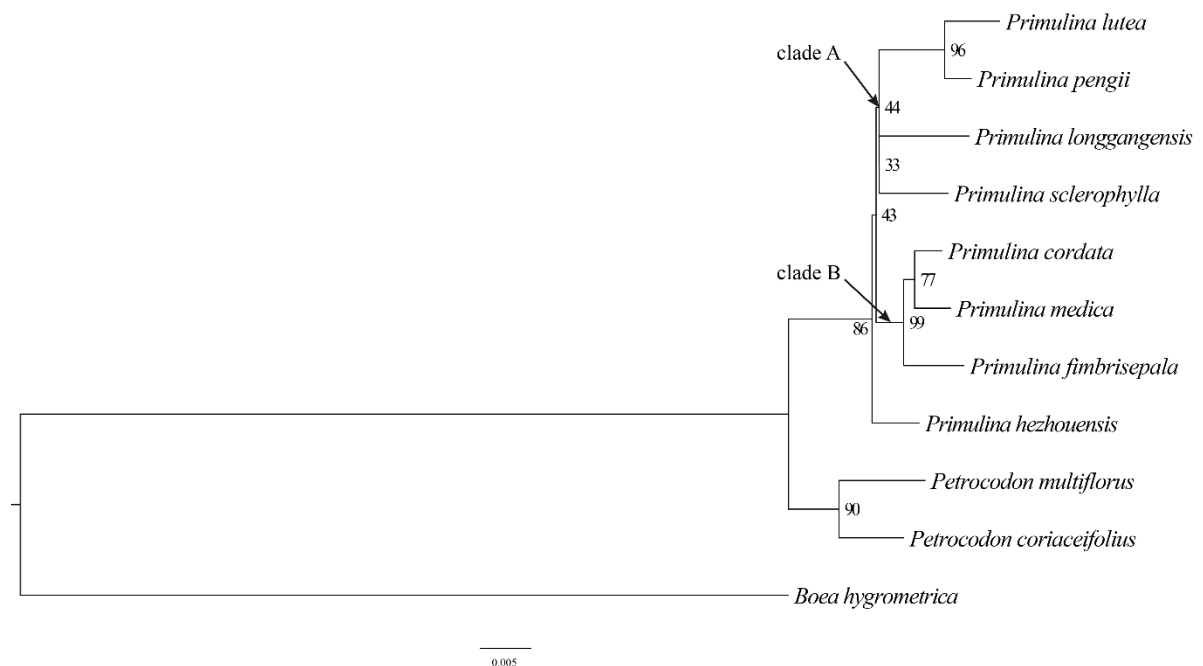
Fig. 13. The ML phylogeny inferred from nr DNA of 11 Gesneriaceae species.
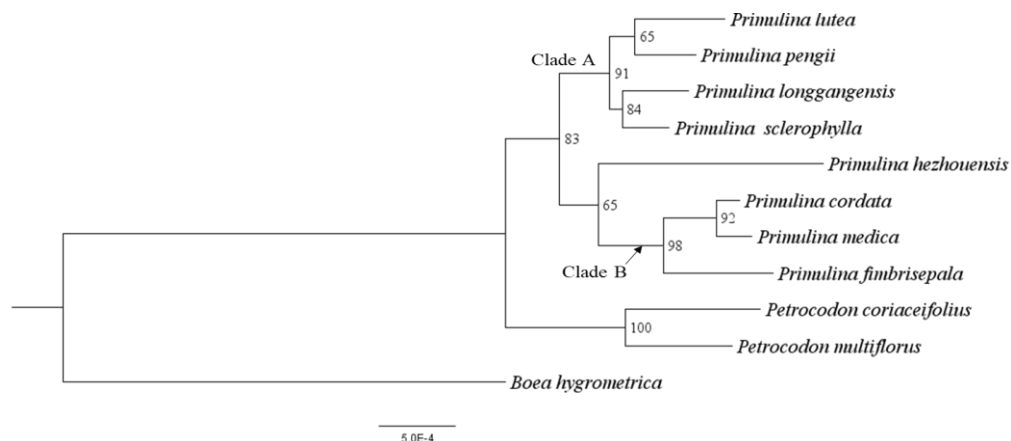Each node was labelled with bootstrap support.

4. Discussion

4.1 Highly conserved *Primulina* and *Petrocodon* plastomes

The comparative analyses of the plastomes of Gesneriaceae species included in this study suggest that these plastomes not only exhibit high level of coding sequence conservation (Fig. 6), but also possess similar genome sizes (Table 3) and highly conserved genome structure without significant structural reconfigurations, such as gene losses or large inversions (Fig. 6). The stable structure of plastomes newly sequenced here may be related to the few number of disperse repeats being identified, since the numbers of repeats in plastomes was shown to correlate with the structural rearrangements in plastomes (Weng *et al.*, 2014). Moreover, the IR/SC boundaries comparison graph also shows that only very minor boundary shifts occurred among the species included in the comparison (Fig. 8). These facts suggest that the overall plastome structure among the Gesneriad plastomes included in the analyses are highly conserved. However, the taxa sampling of this study cannot further confirm if the plastome structure of the entire Gesneriaceae family is consistently highly conserved or not. Perhaps in the foreseeable future, a myriad of plastomes will be sequenced and published as the sequencing technologies continue to advance, so will the plastomes of more Gesneriaceae specie being sequenced, until then can we understand more about the plastome evolution of this family.

4.2 Variable regions identified for *Primulina* and *Petrocodon*

Despite the conservatism of plastome sequences comparing to the nuclear genome (Wolfe *et al.*, 1987), several variable regions among *Primulina* and *Petrocodon* species were identified by the sliding window analysis. Among the variable regions identified by this study, only *rpl32-trnL* intergenic spacer had been employed by previous studies of reconstructing the phylogenies of *Primulina* and *Petrocodon* (Chang, 2012; Ning *et al.*, 2015; Yang *et al.*, 2016). These newly identified variable regions would serve as potential markers for systematics and evolutionary studies about the karst flora.

4.3 Annotation issues in plastome annotations
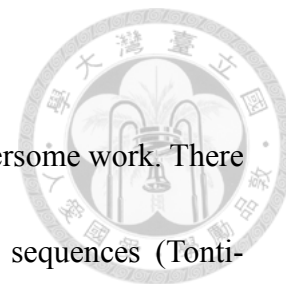
Genome annotation has been considered as a trivial and cumbersome work. There are many erroneous annotations even in the published plastome sequences (Tonti-Filippini *et al.*, 2017). There is no explicit rule about the nomenclature of plastid genes either a universal criteria for gene annotation. Moreover, many of the inadequate annotations were further propagated by new annotations using the previously published yet erroneously annotated ones (Tonti-Filippini *et al.*, 2017). In this study, the annotations of plastome newly sequenced, as well as those accessed from GenBanks, were annotated and checked carefully by GeSeq annotating tool and Geneious.

The initial annotation for my plastid genomes were largely relied on the reference Bhcp genome which may contain some controversial and erroneous annotations. After referring to other published plastomes, I noticed that the *petB* and *petD* genes in Bhcp were both lacking one exon which were extremely short in length of only 6 and 10 bp, respectively. Many of the published plastomes were also lacking of the tiny exons of these two genes. Based on the work of Tanaka *et al.* (1987), I considered the intron containing *petB* and *petD* genes should be the correct annotations. Another case is *rps12* gene, a trans-splicing gene comprising one exon located in LSC, two downstream exons located and duplicated in IRs. The annotation of this gene in other available Gesneriaceae plastomes were lacking of one exon in the duplicated copies in IRs,

rendering the gene lack of stop codon. By referring to the study of *Nicotiana tabacum* chloroplast *rps12* gene (Hildebrand *et al.*, 1988), I annotated this gene containing three exons to all the plastomes included in this study.

In addition to the annotation issues of exon or gene length, there is also disparity in the gene nomenclature among different published data. For instance, the *psbZ* gene in Bhcp genome was annotated as *lhbA* gene which is not a frequently used gene name in most of the plastome annotations. The *psi-psbT* gene, which is more frequently annotated as *psbB* gene, in Bhcp genome was another case of different nomenclature. To be consistent with the general annotations, the annotation of *psbB* gene was adopted in the newly sequenced plastomes.

Regarding to the abovementioned cases, I suggest that the annotation processes should be more carefully curated by referring to general publications. There should be a more explicit plastid gene nomenclature and annotating criteria.

4.4 Incongruences between plastome phylogeny and previous studies

The phylogeny inferred from whole plastome sequences is not completely congruent to the tress generated by previous phylogenetic studies of *Primulina* (Guo *et al.*, 2015; Ning *et al.*, 2015; Yang *et al.*, 2016) which incorporated the combined dataset of ITS and several plastid markers.

One of the differences is the relationships among the three species within clade B of my tree (Fig. 11). My result has resolved *P. medica* as the sister group of *P. fimbrisepala*, and these two species together formed the sister group of *P. corada* and *P. brachytricha* var. *manibracteata*. The phylogenies of Gou *et al.* (2015) and Ning *et al.* (2015), instead, resolved *P. medica* as more close-related to *P. cordata* rather than *P. frimbrisepala* but the former study with weak bootstrap support.

Another incongruence is the phylogenetic position of *P. huaijiensis*. It is the sister species of *P. eburnea,* clustering within clade A in my result (Fig. 11). However, the species was resolved in the clade containing *P. sclerophylla*, *P. liboensis* and *P. hezhouensis* in Guo *et al.* (2015) with very weak bootstrap support. On the contrary, in the phylogenetic trees constructed by Yang *et al.* (2016) and Ning *et al.* (2015), instead of clustering within the clade of *P. sclerophylla* and *P. liboensis*, the species was resolved within the clade containing *P. eburnea, P. lutea, P. longgangensis* and *P.*

linearifolia, the relationship which is congruent to my tree, but with polytomous topolog. Briefly speaking, the phylogenetic position of *P. huaijiensis* was not well resolved in previous studies. Since all of these three studies incorporated the combined dataset of nuclear and plastid data to infer phylogenies, there may be some conflicts between two different genome compartments due to the independent evolutionary histories or stochastic effects. Therefore the bootstrap values are lower than my plastome tree and the topology is not completely congruent with my result. However, even in my tree which is inferred from complete plastomes, the phylogenetic position of *P. huaijiensis* was still not well supported by bootstrap resampling. More efforts should be done to further confirm the indefinite evolutionary position of *P. huaijiensis*.
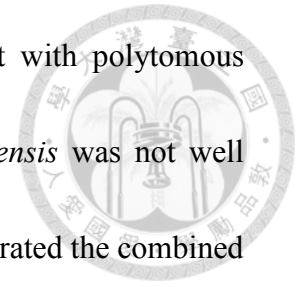
4.5 Incongruences among mitochondrial, nuclear and plastid phylogenies

The topologies of the phylogenies inferred from plastomes, nrDNA and mt exomes are not congruent (Fig. 14, 15, 16). In cp data, *P. hezhouensis* and *P. sclerophylla* are sister species (Fig. 14). However, in mt phylogeny, *P. hezhouensis* clustered with the clade comprising *P. cordata*, *P. medica* and *P. fimbrisepala* (clade A) with relatively long branch length and moderate statistical support, and *P. sclerophylla* clustered with *P. longgangensis* (Fig. 12, 14, 15). On the other hand, in nrDNA phylogeny, *P. hezhouensis* is sister to the rest of the *Primulina* species with weak bootstrap support (Fig.13, 15, 16), and *P. sclerophylla* and *P. longgangensis* are sister species which is consistent with the result in the clade A in mt phylogeny (Fig. 16).

The relationships among the species in clade B of the mt and nuclear phylogenies are congruent (Fig. 16), but differ from the topology of clade B in cp phylogeny (Fig. 14, 15). The overall topology of the mt phylogeny, instead of being congruent with the cp phylogeny, is more similar to the relationships resolved in nuclear tree (Fig. 16). The relationships among the three species of clade B in mt and nuclear trees are in agreement with the result of Guo *et al.* (2015) which was inferred from combined dataset of *trnL-F*, *trnH-psbA* and ITS.

56

Fig. 14. The comparison of ML phylogenetic trees inferred from plastomes (left) and mitochondrial exome (right) data.

Each node was labelled with bootstrap supports.

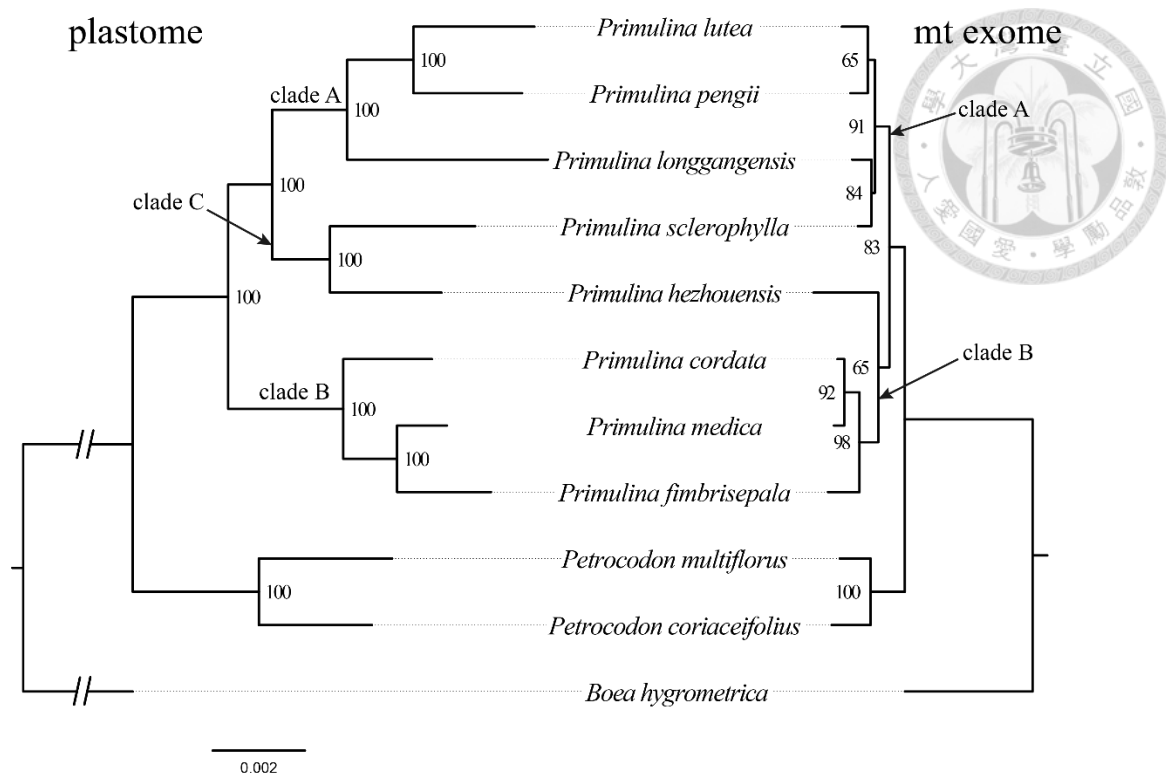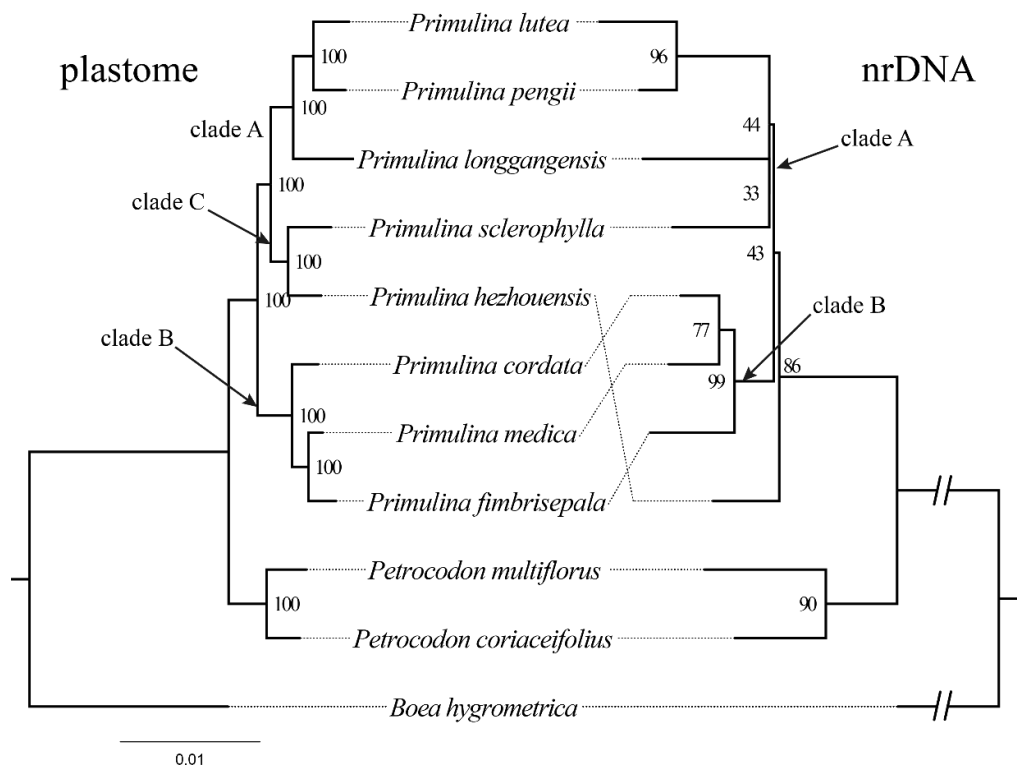**Fig. 15. The comparison of ML phylogenetic trees inferred from plastomes (left) and nrDNA (right) data.**

Each node was labelled with bootstrap supports.

57
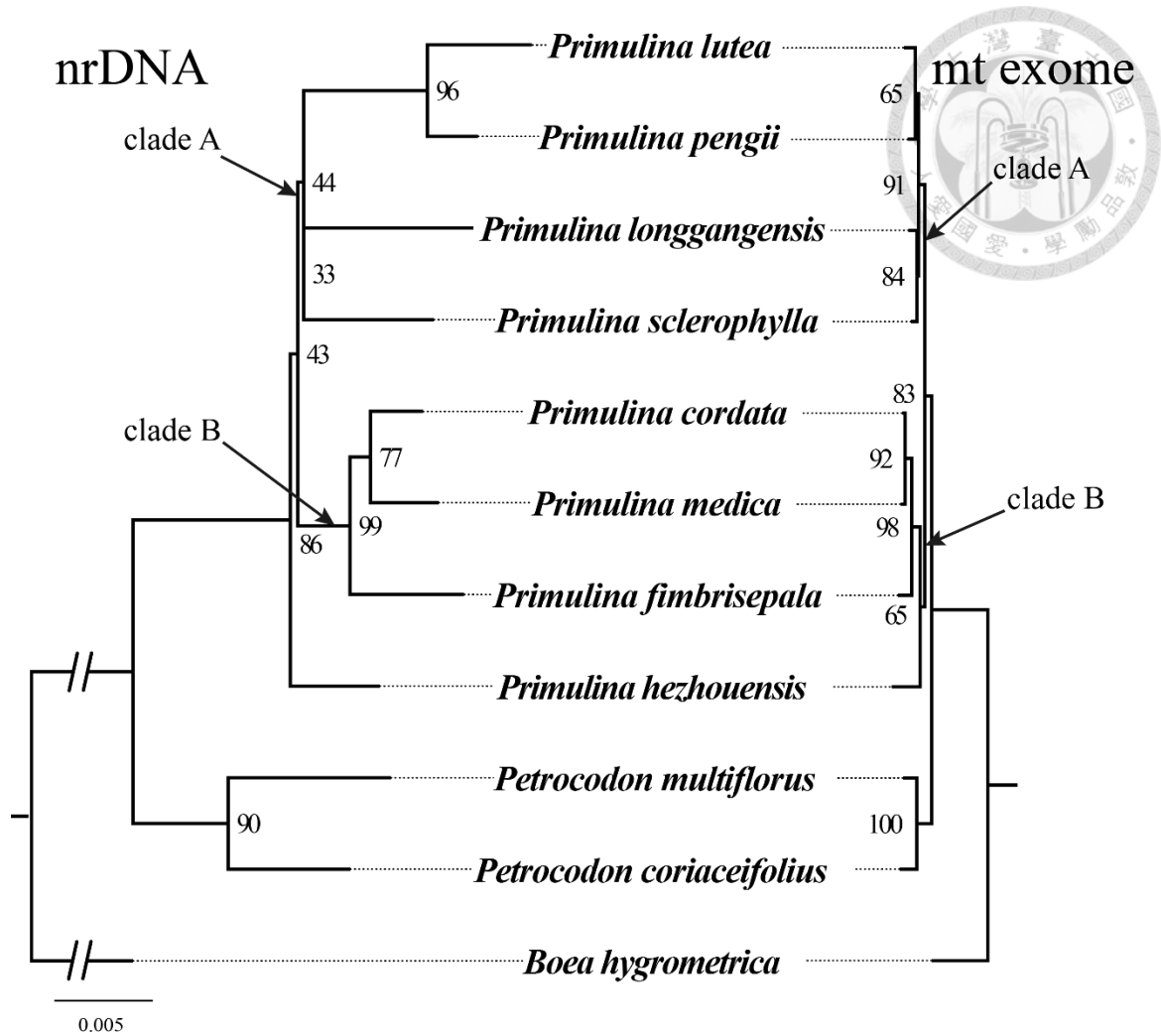
Fig. 16. The comparison of ML phylogenetic trees inferred from nrDNA (left) and mt exome (right) data.

Each node was labelled with bootstrap supports.

4.6 Potential explanations for the incongruences

There may be several potential explanations for the incongruences:

(1) Hybridization

Reconstructing both of the plastid and nuclear phylogenies have been utilized to tackle with the problems of hybridizations or polyploidy which are ubiquitous in many lineages of angiosperms by comparing the different genetic histories they reflected respectively (e.g., Soltis and Kuzoff, 1995; Valcárcel *et al.*, 2003; Govindarajulu *et al.*, 2011; Krak *et al.*, 2012). Therefore, the phylogenetic incongruences, especially between nuclear and organellar phylogenies, may be the results from hybridization or polyploidy. A few hybridization events in the genus *Primulina* has been recorded (Xu *et al.*, 2013; Zhang *et al.*, 2017). Earlier studies suggested that there is only one polyploidy level observed in *Primulina* whose chromosome number is 2n=36 (Christie *et al.*, 2012; Liu *et al.*, 2012; Kang *et al.*, 2014). Hence, the incongruences among nuclear and plastome phylogenies may due to the hybridization events in *Primulina*. The species which did not have consistent relationship among the cp, mt and nuclear phylogenies, such as *P. hezhouensis*, should be observed with scrutiny to further confirm if there is hybridization in this genus.

(2) Paternal leakage of mitochondria
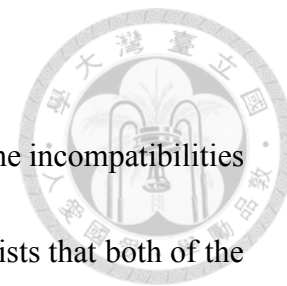
The hybridization alone may not plausible enough to explain the incompatibilities between mt and cp phylogenies. It is commonly held among biologists that both of the organelles are maternally inherited, and the phylogenies based on both data are supposed to represent the same maternal history in many plants (Rieseberg and Soltis, 1991; Olson and McCauley, 2000). However, other studies have reported the biparental transmission of cytoplasmic genomes (e.g., Nagata *et al.*, 1999; Zhang *et al.*, 2003; McCauley *et al.*, 2005). As mentioned before, low level of paternally inherited plastids has also been recorded (Reboud and Zeyl, 1994; Thyssen *et al.*, 2012). Moreover, there is increasing evidence of paternally transmitted mitochondrial DNA via pollen, namely, 'paternal leakage' (Havey, 1997; Ellis *et al.*, 2008; Pearl *et al.*, 2009; McCauley, 2013). These findings make the assumption of the consistence of two organellar phylogenies to be questioned. On the other hand, these facts can also be considered as additional information, providing complementary insights into different evolutionary histories of the organisms. In my results, considering the phylogenies inferred from mt and nuclear data are more similar to each other than to the cp tree (Fig. 14, 15, 16), paternal leakage (PL) may be the most feasible explanation for the incongruences. PL is usually observed in gynodioecious plants, in which the cytoplasmic male sterility (CMS) may facilitate the paternal mitochondria transmission (Tsitrone *et al.*, 2003; McCauley *et al.*, 2005;

60

Mandel *et al.*, 2012; McCauley, 2013). However, to my knowledge, there is no gynodioecious species in the hermaphroditic genus *Primulina*. Perhaps more field observations are needed to explore the sexual reproductive system of the genus and to better explain the result of cytoplasmic incongruences.
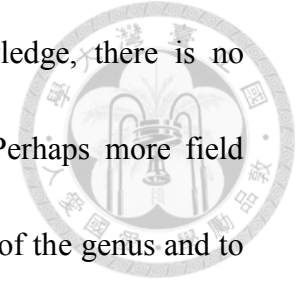
The biparental transmission of mitochondria, resulting mt heteroplasmy, will form intraindividual mtDNA diversity which is thought to exert the selective pressures on nuclear genome to restore the homogeneous cytoplasm, limiting the selfish CMS of mt genes to propagate (McCauley, 2013). This 'cyto-nuclear dynamic' will lead to the rapid fixation of CMS mutants (Bailey and Delph, 2007), and this male sterility process can facilitate chloroplast capture which is another possible causation of phylogenetic incongruences (Tsitrone *et al.*, 2003).

(3) Processed paralogs in mt genes

The organellar incongruences might also result from the incorporation of 'processed paralogs' after RNA editing in the mt data. Several studies have investigated the influence of processed paralogs on phylogenetic reconstructions, showing that the edited genes cause the incongruence between different genome compartments (Bowe and dePamphilis, 1996; Petersen *et al.*, 2006a). The incorporation of processed paralogs may introduce the erroneous signals in the phylogenetic analyses.

(4) Rate heterogeneity of mt genes among different lineages
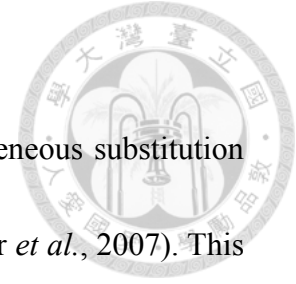
Another possible causation may be the diverged and heterogeneous substitution rate among different lineages of mt genes (Cho *et al.*, 2004; Mower *et al.*, 2007). This can potentially lead to long-branch attraction (Petersen *et al.*, 2006a; Petersen *et al.*, 2006b). In my mt phylogeny, the branch length of *P. hezhouensis* is relatively long comparing to other taxa (Fig. 11). Referring to the relationships based on plastome data, the species is the sister group of *P. sclerophylla* (Fig. 10). On the other hand, in mt phylogeny, *P. sclerophylla* possesses very short branch length which may cause *P. hezhouensis* not clustering with *P. sclerophylla*, forming the incongruence between two phylogenies.

(5) Recombination, gene transfer and gene conversions occurred in mt genomes

Lastly, the horizontal gene transfer mediated by genome recombination, mitochondria fusion and/or gene conversion in the mitochondrial genomes have been documented to cause the incongruences as well (Hao *et al.*, 2010; Rice *et al.*, 2013). Therefore, different mt genes may reflect different genetic histories. The concatenation of every coding exons, which may contain homoplasy, may result in problematic relationships.

4.7 Utilities of mitochondrial data in plant phylogenetic analyses
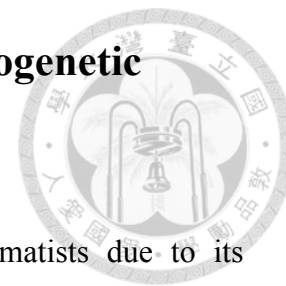
The plant mitochondrial data is not popular among systematists due to its recombining and dynamic structure (Galtier, 2011), extensive RNA editing with processed paralogs (Petersen *et al.*, 2006a), and three time slower evolutionary rate comparing to plastomes (Wolfe *et al.*, 1987). These features have made plant mitochondria sequences problematic and uninformative for phylogenetic studies.

Although a total of 26 mitochondrial protein coding genes have been recovered and concatenated, there are not much informative sites (1.8%) due to the extremely conserve nature of mt coding sequences, resulting the lower resolution of phylogeny comparing to cp data (Fig. 14). The phylogeny based on this data is also incongruent with the plastid ones (Fig. 14) and the possible causations were discussed above. Despite the potentially problematic data, the mitochondria sequences still serve as another source of information. It may represent a different genetic history from cp or nuclear genome, providing potentials to disentangle complex speciation processes of plants (Govindarajulu *et al.*, 2011). By using genome skimming method, which can easily acquire the mitochondrial data, and benefitting from the rapidly advancing sequencing technologies, researchers are able to explore this complex but underutilized part of genomes, probing into deeper evolutionary questions.

5. Conclusion

By employing the genome skimming strategy, the complete plastomes of eight *Primulina* and two *Petrocodon* species were newly sequenced and assembled. Together with other published Gesneriaceae plastomes, no large inversions, gene losses or pseudogenes have been detected. The IR/SC boundary shifts are very minor as well. The plastome structures of *Primulina*, *Petrocodon* and three other Gesneriaceae species included in the analyses are highly conserved.

A total of 18 complete plastome sequences of Gesneriaceae species have been employed to reconstruct a statistically highly supported phylogenetic tree representing a maternal evolutionary history of *Primulina* and its related genera. This can be regarded as a back bone relationship for the two karst characteristic genera, especially for *Primulina* which contains diverse species. Besides assembling and analyzing the sequences of plastid genomes, mitochondrial exomes and nuclear ribosomal cistrons have been recovered as well. Although the phylogeny inferred from mt data is incongruent with cp tree, it provides a first glimpse to the mitochondrial history of *Primulina* and *Petrocodon*. Together with phylogeny inferred from nrDNA, the genome skimming data offer the chance to compare three different parts of genomic DNAs of these two genera. The phylogenetic incongruences have also demonstrated the

potentials of hybridization and paternal leakage of mitochondria, providing novel

insights into the speciation processes of the genus *Primulina*.
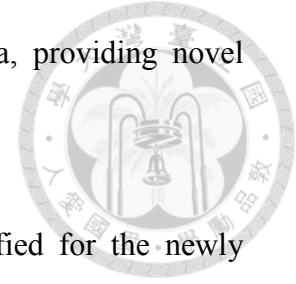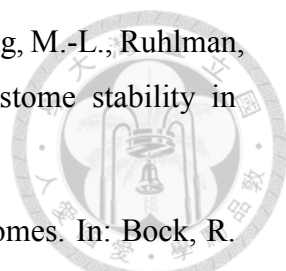
A number of dispersed repeats and SSRs were also identified for the newly

sequenced species. Lastly, I identified eight variable regions, which can serve as

potential useful markers for further phylogenetic studies of these limestone genera. The

data would improve our understandings of the evolution of karst flora.

6. References
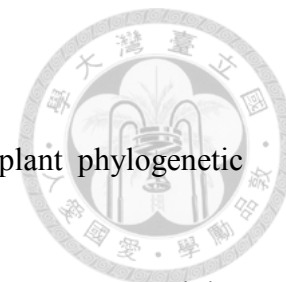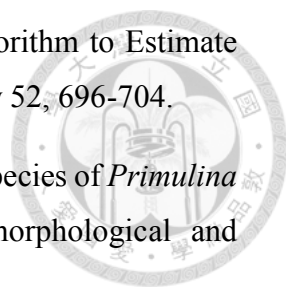
Álvarez, I., Wendel, J.F., 2003. Ribosomal ITS sequences and plant phylogenetic inference. Molecular Phylogenetics and Evolution 29, 417-434.

Alverson, A.J., Wei, X., Rice, D.W., Stern, D.B., Barry, K., Palmer, J.D., 2010. Insights into the Evolution of Mitochondrial Genome Size from Complete Sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). Molecular Biology and Evolution 27, 1436-1448.

Andrews, S., 2010. FastQC A Quality Control tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Appels, R., Honeycutt, R.L., 1986. rDNA : evolution over a billion years. In, DNA systematic : Plants. CRC Press Inc., Boca Raton, Florida, pp. 82-135.

Bailey, M.F., Delph, L.F., 2007. A field guide to models of sex-ratio evolution in gynodioecious species. Oikos 116, 1609-1617.

Barber, J.C., Finch, C.C., Francisco-Ortega, J., Santos-Guerra, A., Jansen, R.K., 2007. Hybridization in Macaronesian *Sideritis* (Lamiaceae): evidence from incongruence of multiple independent nuclear and chloroplast sequence datasets. Taxon 56, 74-88.

Beier, S., Thiel, T., Münch, T., Scholz, U., Mascher, M., 2017. MISA-web: a web server for microsatellite prediction. Bioinformatics 33, 2583-2585.

Bendich, A.J., 1987. Why do chloroplasts and mitochondria contain so many copies of their genome? BioEssays 6, 279-282.

Bendich, A.J., 2004. Circular Chloroplast Chromosomes: The Grand Illusion. The Plant Cell 16, 1661-1666.

Birky, C.W., 1995. Uniparental inheritance of mitochondrial and chloroplast genes: mechanisms and evolution. Proceedings of the National Academy of Sciences 92, 11331-11338.

Birky, C.W., Walsh, J.B., 1992. Biased gene conversion, copy number, and apparent mutation rate differences within chloroplast and bacterial genomes. Genetics 130, 677-683.

Blazier, J.C., Jansen, R.K., Mower, J.P., Govindu, M., Zhang, J., Weng, M.-L., Ruhlman, T.A., 2016. Variable presence of the inverted repeat and plastome stability in *Erodium*. Ann. Bot. 117, 1209-1220.

Bock, R., 2007. Structure, function, and inheritance of plastid genomes. In: Bock, R. (Ed.), Cell and Molecular Biology of Plastids. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 29-63.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114-2120.

Bowe, L.M., dePamphilis, C.W., 1996. Effects of RNA editing and gene processing on phylogenetic reconstruction. Molecular Biology and Evolution 13, 1159-1166.

Bryan, G.J., McNicoll, J., Ramsay, G., Meyer, R.C., De Jong, W.S., 1999. Polymorphic simple sequence repeat markers in chloroplast genomes of Solanaceous plants. Theoretical and Applied Genetics 99, 859-867.

Cato, S.A., Richardson, T.E., 1996. Inter- and intraspecific polymorphism at chloroplast SSR loci and the inheritance of plastids in *Pinus radiata* D. Don. Theoretical and Applied Genetics 93, 587-592.

Cen, H.-F., Fu, L.-F., Wen, F., 2017. *Petrocodon urceolatus* sp. nov. (Gesneriaceae) from Zhangjiajie city, northwestern Hunan Province, China. Nordic Journal of Botany 35, 334-338.

Chang, H., 2012. Tempo and mode of plant evolution on limestone karst in Guangxi, China—Insights from phylogenetics of *Primulina* (Gesneriaceae). In, School of Forestry and Resource Conservation. National Taiwan University.

Chen, W.-H., Möller, M., Shui, Y.-M., Wang, H., Yang, J.-B., Li, G.-Y., 2014. Three new species of *Petrocodon* (Gesneriaceae), endemic to the limestone areas of southwest China, and preliminary insights into the diversification patterns of the genus. Systematic Botany 39, 316-330.

Cho, Y., Mower, J.P., Qiu, Y.-L., Palmer, J.D., 2004. Mitochondrial substitution rates are extraordinarily elevated and variable in a genus of flowering plants. Proc. Natl. Acad. Sci. U. S. A. 101, 17741-17746.

67

Christie, F., Barber, S., Möller, M., 2012. New chromosome counts in Old World Gesneriaceae: numbers for species hitherto regarded as *Chirita*, and their systematic and evolutionary significance. Edinburgh Journal of Botany 69, 323-345.

Chumley, T.W., Palmer, J.D., Mower, J.P., Fourcade, H.M., Calie, P.J., Boore, J.L., Jansen, R.K., 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, 2175-2190.

Chung, K.-F., Huang, H.-Y., Peng, C.-I., Xu, W.-B., 2013. *Primulina mabaensis* (Gesneriaceae), a new species from a limestone cave of northern Guangdong, China. Phytotaxa 92, 40-48.

Corriveau, J.L., Coleman, A.W., 1988. Rapid Screening Method to Detect Potential Biparental Inheritance of Plastid DNA and Results for Over 200 Angiosperm Species. American Journal of Botany 75, 1443-1458.

Cronn, R., Knaus, B.J., Liston, A., Maughan, P.J., Parks, M., Syring, J.V., Udall, J., 2012. Targeted enrichment strategies for next-generation plant biology. American Journal of Botany 99, 291-311.

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

Darriba, D., Taboada, G.L., Doallo, R., Posada, D., 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat Meth 9, 772-772.

Ding, P., Shao, Y., Li, Q., Gao, J., Zhang, R., Lai, X., Wang, D., Zhang, H., 2016. The complete chloroplast genome sequence of the medicinal plant *Andrographis paniculata*. Mitochondrial DNA Part A 27, 2347-2348.

Doyle, J.J., Doyle, J.L., 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemical Bulletin 19, 11-15.

Doyle, J.J., Doyle, J.L., 1990. Isolation of plant DNA from fresh tissue. Focus 12, 13-15.

El-Metwally, S., Ouda, O.M., Helmy, M., 2014. Next-Generation Sequence Assembly

Overview. In, Next Generation Sequencing Technologies and Challenges in Sequence Assembly. Springer New York, New York, NY, pp. 73-78.

Ellis, J.R., Bentley, K.E., McCauley, D.E., 2008. Detection of rare paternal chloroplast inheritance in controlled crosses of the endangered sunflower *Helianthus verticillatus*. Heredity 100, 574.

Feliner, G.N., Rosselló, J.A., 2007. Better the devil you know? Guidelines for insightful utilization of nrDNA ITS in species-level evolutionary studies in plants. Molecular Phylogenetics and Evolution 44, 911-919.

Feschotte, C., Jiang, N., Wessler, S.R., 2002. Plant transposable elements: where genetics meets genomics. Nature Reviews Genetics 3, 329.

Frazer, K.A., Pachter, L., Poliakov, A., Rubin, E.M., Dubchak, I., 2004. VISTA: computational tools for comparative genomics. Nucleic Acids Res 32, W273-W279.

Galtier, N., 2011. The intriguing evolutionary dynamics of plant mitochondrial DNA. BMC Biol. 9, 61.

Goremykin, V.V., Holland, B., Hirsch-Ernst, K.I., Hellwig, F.H., 2005. Analysis of *Acorus calamus* chloroplast genome and its phylogenetic implications. Molecular Biology and Evolution 22, 1813-1822.

Govindarajulu, R., Hughes, C.E., Alexander, P.J., Bailey, C.D., 2011. The complex evolutionary dynamics of ancient and recent polyploidy in *Leucaena* (Leguminosae; Mimosoideae). American Journal of Botany 98, 2064-2076.

Graham, S.W., Reeves, P.A., Burns, A.C.E., Olmstead, R.G., 2000. Microstructural Changes in Noncoding Chloroplast DNA: Interpretation, Evolution, and Utility of Indels and Inversions in Basal Angiosperm Phylogenetic Inference. Int J Plant Sci 161, S83-S96.

Gray, M.W., Archibald, J.M., 2012. Origins of Mitochondria and Plastids. In: Bock, R., Knoop, V. (Eds.), Genomics of Chloroplasts and Mitochondria. Springer Netherlands, Dordrecht, pp. 1-30.

Greiner, S., Sobanski, J., Bock, R., 2015. Why are most organelle genomes transmitted maternally? BioEssays 37, 80-94.

Guindon, S., Gascuel, O., 2003. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. Systematic Biology 52, 696-704.

Guo, J., Pan, B., Liu, J., Xu, W.-B., Chung, K.-F., 2015. Three new species of *Primulina* (Gesneriaceae) from limestone karsts of China based on morphological and molecular evidence. Botanical Studies 56, 34.

Hagemann, R., 2004. The Sexual Inheritance of Plant Organelles. In: Daniell, H., Chase, C. (Eds.), Molecular Biology and Biotechnology of Plant Organelles: Chloroplasts and Mitochondria. Springer Netherlands, Dordrecht, pp. 93-113.

Hance, H.F., 1883. New Chinese *Cytrandreæ*. Journal of Botany, British and foreign 21, 169.

Hao, W., Richardson, A.O., Zheng, Y., Palmer, J.D., 2010. Gorgeous mosaic of mitochondrial genes created by horizontal transfer and gene conversion. Proceedings of the National Academy of Sciences 107, 21576-21581.

Havey, M.J., 1997. Predominant Paternal Transmission of the Mitochondrial Genome in Cucumber. Journal of Heredity 88, 232-235.

Henriquez, C.L., Arias, T., Pires, J.C., Croat, T.B., Schaal, B.A., 2014. Phylogenomics of the plant family Araceae. Mol Phylogenet Evol 75, 91-102.

Hildebrand, M., Hallick, R.B., Passavant, C.W., Bourque, D.P., 1988. Trans-splicing in chloroplasts: the *rps12* loci of *Nicotiana tabacum*. Proceedings of the National Academy of Sciences 85, 372-376.

Hou, M.-F., López-Pujol, J., Qin, H.-N., Wang, L.-S., Liu, Y., 2010. Distribution pattern and conservation priorities for vascular plants in Southern China: Guangxi Province as a case study. Botanical Studies 51, 377-386.

Hupfer, H., Swiatek, M., Hornung, S., Herrmann, R.G., Maier, R.M., Chiu, W.-L., Sears, B., 2000. Complete nucleotide sequence of the *Oenothera elata* plastid chromosome, representing plastome I of the five distinguishable Euoenothera plastomes. Molecular and General Genetics MGG 263, 581-585.

Isoda, K., Shiraishi, S., Watanabe, S., Kitamura, K., 2000. Molecular evidence of natural hybridization between *Abies veitchii* and *A. homolepis* (Pinaceae) revealed

by chloroplast, mitochondrial and nuclear DNA markers. Molecular Ecology 9, 1965-1974.

Jiang, D., Zhao, Z., Zhang, T., Zhong, W., Liu, C., Yuan, Q., Huang, L., 2017. The Chloroplast Genome Sequence of Scutellaria baicalensis Provides Insight into Intraspecific and Interspecific Chloroplast Genome Diversity in Scutellaria. Genes 8, 227.

Kang, M., Tao, J., Wang, J., Ren, C., Qi, Q., Xiang, Q.-Y., Huang, H., 2014. Adaptive and nonadaptive genome size evolution in Karst endemic flora of China. New Phytologist 202, 1371-1381.

Kato, T., Kaneko, T., Sato, S., Nakamura, Y., Tabata, S., 2000. Complete structure of the chloroplast genome of a legume, *Lotus japonicus*. DNA research : an international journal for rapid publication of reports on genes and genomes 7, 323-330.

Katoh, K., Misawa, K., Kuma, K., Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30, 3059-3066.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., Drummond, A., 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28, 1647-1649.

Kelchner, S.A., 2000. The Evolution of Non-Coding Chloroplast DNA and Its Application in Plant Systematics. Ann Mo Bot Gard 87, 482-498.

Knox, E.B., Palmer, J.D., 1995. Chloroplast DNA variation and the recent radiation of the giant senecios (Asteraceae) on the tall mountains of eastern Africa. Proceedings of the National Academy of Sciences 92, 10349-10353.

Kolodner, R., Tewari, K.K., 1979. Inverted repeats in chloroplast DNA from higher plants. Proceedings of the National Academy of Sciences 76, 41-45.

Krak, K., Caklová, P., Chrtek, J., Fehrer, J., 2012. Reconstruction of phylogenetic

relationships in a highly reticulate group with deep coalescence and recent speciation (*Hieracium*, Asteraceae). Heredity 110, 138.

Ku, C., Hu, J.-M., Kuo, C.-H., 2013. Complete Plastid Genome Sequence of the Basal Asterid *Ardisia polysticta* Miq. and Comparative Analyses of Asterid Plastid Genomes. PLOS ONE 8, e62548.

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 Res 29, 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 6, 21669.

Li, J.-M., Wang, Y.-Z., 2007. Phylogenetic reconstruction among species of *Chiritopsis* and *Chirita* sect. *Gibbosaccus* (Gesneriaceae) based on nrDNA ITS and cpDNA *trnL-F* sequences. Systematic Botany 32, 888-898.

Librado, P., Rozas, J., 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25, 1451-1452.

Lilly, J.W., Havey, M.J., Jackson, S.A., Jiang, J., 2001. Cytogenomic Analyses Reveal the Structural Plasticity of the Chloroplast Genome in Higher Plants. The Plant Cell 13, 245-254.

Lin, C.-P., Huang, J.-P., Wu, C.-S., Hsu, C.-Y., Chaw, S.-M., 2010. Comparative Chloroplast Genomics Reveals the Evolution of Pinaceae Genera and Subfamilies. Genome Biol Evol 2, 504-517.

Liu, R., Pan, B., Zhou, T., Liao, J., 2012. Cytological studies on *Primulina* taxa (Gesneriaceae) from limestone karsts in Guangxi province, China. Caryologia 65, 295-303.

Lohse, M., Drechsel, O., Kahlau, S., Bock, R., 2013. OrganellarGenomeDRAW—a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. Nucleic Acids Res 41, W575-W581.

Lowe, T.M., Chan, P.P., 2016. tRNAscan-SE On-line: integrating search and context

for analysis of transfer RNA genes. Nucleic Acids Res 44, W54-W57.

Lowe, T.M., Eddy, S.R., 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25, 955-964.

Lu, Y.-B., Pan, B., Huang, J., Xiang, C.-L., Zhang, Q., 2017. *Petrocodon pulchriflorus* sp. nov. (Gesneriaceae) from Guangxi, China. 2017 291, 8.

Möller, M., Forrest, A., Wei, Y.-G., Weber, A., 2011. A molecular phylogenetic assessment of the advanced Asiatic and Malesian Didymocarpoid Gesneriaceae with focus on non-monophyletic and monotypic genera. Plant Syst. Evol. 292, 223-248.

Möller, M., Pfosser, M., Jang, C.G., Mayer, V., Clark, A., Hollingsworth, M.L., Barfuss, M.H.J., Wang, Y.Z., Kiehn, M., Weber, A., 2009. A preliminary phylogeny of the 'Didymocarpoid Gesneriaceae' based on three molecular data sets: incongruence with available tribal classifications. American Journal of Botany 96, 989-1010.

Möller, M., Wei, Y.G., Wen, F., Clark, J.L., Weber, A., 2016. You win some you lose some: updated generic delineations and classification of Gesneriaceae–implications for the family in China. Guihaia 36, 44-60.

Malé, P.-J.G., Bardon, L., Besnard, G., Coissac, E., Delsuc, F., Engel, J., Lhuillier, E., Scotti-Saintagne, C., Tinaut, A., Chave, J., 2014. Genome skimming by shotgun sequencing helps resolve the phylogeny of a pantropical tree family. Molecular Ecology Resources 14, 966-975.

Mandel, J.R., McAssey, E.V., Roland, K.M., McCauley, D.E., 2012. Mitochondrial Gene Diversity Associated with the *atp9* Stop Codon in Natural Populations of Wild Carrot (*Daucus carota* ssp. *carota*). Journal of Heredity 103, 418-425.

Martin, G., Baurens, F.-C., Cardi, C., Aury, J.-M., D'Hont, A., 2013. The Complete Chloroplast Genome of Banana (*Musa acuminata*, Zingiberales): Insight into Plastid Monocotyledon Evolution. PLOS ONE 8, e67350.

Martin, W., Rujan, T., Richly, E., Hansen, A., Cornelsen, S., Lins, T., Leister, D., Stoebe, B., Hasegawa, M., Penny, D., 2002. Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proceedings of the National Academy of

Sciences 99, 12246-12251.

McCauley, D.E., 2013. Paternal leakage, heteroplasmy, and the evolution of plant mitochondrial genomes. New Phytologist 200, 966-977.

McCauley, D.E., Bailey, M.F., Sherman, N.A., Darnell, M.Z., 2005. Evidence for paternal transmission and heteroplasmy in the mitochondrial genome of *Silene vulgaris*, a gynodioecious plant. Heredity 95, 50.

Metzlaff, M., Börner, T., Hagemann, R., 1981. Variations of chloroplast DNAs in the genus *Pelargonium* and their biparental inheritance. Theoretical and Applied Genetics 60, 37-41.

Meyer, B., Stubbe, W., 1974. Das Zahlenverhältnis von mütterlichen und väterlichen Piastiden in den Zygoten von *Oenothera erythrosepala* Borbas (syn. *Oe. lamarckiana*). Berichte der Deutschen Botanischen Gesellschaft 87, 29-38.

Meyers, B.C., Tingey, S.V., Morgante, M., 2001. Abundance, Distribution, and Transcriptional Activity of Repetitive Elements in the Maize Genome. Genome Research 11, 1660-1676.

Mogensen, H.L., 1996. The Hows and Whys of Cytoplasmic Inheritance in Seed Plants. American Journal of Botany 83, 383-404.

Mower, J.P., Sloan, D.B., Alverson, A.J., 2012. Plant Mitochondrial Genome Diversity: The Genomics Revolution. In: Wendel, F.J., Greilhuber, J., Dolezel, J., Leitch, J.I. (Eds.), Plant Genome Diversity Volume 1: Plant Genomes, their Residents, and their Evolutionary Dynamics. Springer Vienna, Vienna, pp. 123-144.

Mower, J.P., Touzet, P., Gummow, J.S., Delph, L.F., Palmer, J.D., 2007. Extensive variation in synonymous substitution rates in mitochondrial genes of seed plants. BMC Evol Biol 7, 135.

Nagata, N., Saito, C., Sakai, A., Kuroiwa, H., Kuroiwa, T., 1999. The selective increase or decrease of organellar DNA in generative cells just after pollen mitosis one controls cytoplasmic inheritance. Planta 209, 53-65.

Ni, L., Zhao, Z., Dorje, G., Ma, M., 2016. The Complete Chloroplast Genome of Ye-Xing-Ba (*Scrophularia dentata*; Scrophulariaceae), an Alpine Tibetan Herb. PLOS

ONE 11, e0158488.
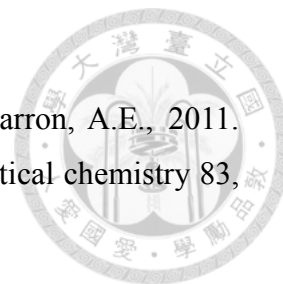
Niedringhaus, T.P., Milanova, D., Kerby, M.B., Snyder, M.P., Barron, A.E., 2011. Landscape of Next-Generation Sequencing Technologies. Analytical chemistry 83, 4327-4341.

Ning, Z.L., LI, D.M., Pan, B., Kang, M., 2015. Primulina rubribracteata, a new species of Gesneriaceae from southern Hunan, China.

Oldenburg, D.J., Bendich, A.J., 2004. Changes in the Structure of DNA Molecules and the Amount of DNA Per Plastid During Chloroplast Development in Maize. Journal of Molecular Biology 344, 1311-1330.

Olson, M.S., McCauley, D.E., 2000. Linkage disequilibrium and phylogenetic congruence between chloroplast and mitochondrial haplotypes in *Silene vulgaris*. Proceedings of the Royal Society of London. Series B: Biological Sciences 267, 1801-1808.

Palmer, J.D., 1983. Chloroplast DNA exists in two orientations. Nature 301, 92.

Palmer, J.D., 1985. Comparative Organization of Chloroplast Genomes. Annual Review of Genetics 19, 325-354.
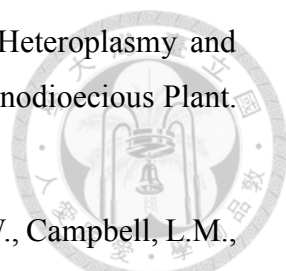
Palmer, J.D., 1991. Plastid chromosomes: structure and evolution. In: Bogorad, L., Vasil, I.K. (Eds.), The molecular biology of plastids. Academic Press, New York, pp. 5-53.

Palmer, J.D., Jansen, R.K., Michaels, H.J., Chase, M.W., Manhart, J.R., 1988. Chloroplast DNA Variation and Plant Phylogeny. Ann Mo Bot Gard 75, 1180-1206.

Palmer, J.D., Stein, D.B., 1986. Conservation of chloroplast genome structure among vascular plants. Current Genetics 10, 823-833.

Pan, B., Wen, F., Zhao, B., Deng, T., Xu, W.-B., Huang, S.-X., 2013. *Primulina beiliuensis* B. Pan & S. X. Huang, a new species of Gesneriaceae from limeston areas in Guangxi, China. Guihaia 33, 591-598.

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

Pearl, S.A., Welch, M.E., McCauley, D.E., 2009. Mitochondrial Heteroplasmy and Paternal Leakage in Natural Populations of *Silene vulgaris*, a Gynodioecious Plant. Molecular Biology and Evolution 26, 537-545.
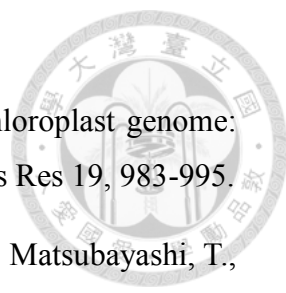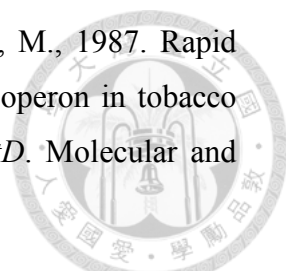
Petersen, G., Seberg, O., Davis, J.I., Goldman, D.H., Stevenson, D.W., Campbell, L.M., Michelangeli, F.A., Specht, C.D., Chase, M.W., Fay, M.F., 2006a. Mitochondrial data in monocot phylogenetics. Aliso 22, 52-62.

Petersen, G., Seberg, O., Davis, J.I., Stevenson, D.W., 2006b. RNA editing and phylogenetic reconstruction in two monocot mitochondrial genes. Taxon 55, 871-886.

Powell, W., Morgante, M., Andre, C., McNicol, J.W., Machray, G.C., Doyle, J.J., Tingey, S.V., Rafalski, J.A., 1995. Hypervariable microsatellites provide a general source of polymorphic DNA markers for the chloroplast genome. Current Biology 5, 1023-1029.

Provan, J., Powell, W., Hollingsworth, P.M., 2001. Chloroplast microsatellites: new tools for studies in plant ecology and evolution. Trends in ecology & evolution 16, 142-147.

Rauwolf, U., Golczyk, H., Greiner, S., Herrmann, R.G., 2009. Variable amounts of DNA related to the size of chloroplasts III. Biochemical determinations of DNA amounts per organelle. Molecular Genetics and Genomics 283, 35.

Reboud, X., Zeyl, C., 1994. Organelle inheritance in plants. Heredity 72, 132.

Rice, D.W., Alverson, A.J., Richardson, A.O., Young, G.J., Sanchez-Puerta, M.V., Munzinger, J., Barry, K., Boore, J.L., Zhang, Y., dePamphilis, C.W., Knox, E.B., Palmer, J.D., 2013. Horizontal Transfer of Entire Genomes via Mitochondrial Fusion in the Angiosperm *Amborella*. Science 342, 1468-1473.

Rieseberg, L.H., Soltis, D., 1991. Phylogenetic consequences of cytoplasmic gene flow in plants. Evolutionary Trends in Plants.

Rochaix, J.-D., 1997. Chloroplast reverse genetics: new insights into the function of plastid genes. Trends in Plant Science 2, 419-425.

Rogers, S.O., Bendich, A.J., 1987. Ribosomal RNA genes in plants: variability in copy

number and in the intergenic spacer. Plant Molecular Biology 9, 509-520.

Rowan, B.A., Oldenburg, D.J., Bendich, A.J., 2009. A multiple-method approach reveals a declining amount of chloroplast DNA during development in *Arabidopsis*. BMC Plant Biology 9, 3.

Ruhfel, B.R., Gitzendanner, M.A., Soltis, P.S., Soltis, D.E., Burleigh, J.G., 2014. From algae to angiosperms–inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. BMC Evol Biol 14, 23.

Sang, T., Crawford, D.J., Stuessy, T.F., 1997. Chloroplast DNA phylogeny, reticulate evolution, and biogeography of *Paeonia* (Paeoniaceae). American Journal of Botany 84, 1120-1136.

Schmidt, T., Heslop-Harrison, J.S., 1998. Genomes, genes and junk: the large-scale organization of plant chromosomes. Trends in Plant Science 3, 195-199.

Schmitz-Linneweber, C., Maier, R.M., Alcaraz, J.-P., Cottet, A., Herrmann, R.G., Mache, R., 2001. The plastid chromosome of spinach (*Spinacia oleracea*): complete nucleotide sequence and gene organization. Plant Molecular Biology 45, 307-315.

Shaver, J.M., Oldenburg, D.J., Bendich, A.J., 2008. The structure of chloroplast DNA molecules and the effects of light on the amount of chloroplast DNA during development in *Medicago truncatula*. Plant Physiology 146, 1064-1074.

Shaw, J., Lickey, E.B., Beck, J.T., Farmer, S.B., Liu, W., Miller, J., Siripun, K.C., Winder, C.T., Schilling, E.E., Small, R.L., 2005. The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. American Journal of Botany 92, 142-166.

Shaw, J., Lickey, E.B., Schilling, E.E., Small, R.L., 2007. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. American Journal of Botany 94, 275-288.

Shaw, J., Shafer, H.L., Leonard, O.R., Kovach, M.J., Schorr, M., Morris, A.B., 2014. Chloroplast DNA sequence utility for the lowest phylogenetic and phylogeographic inferences in angiosperms: The tortoise and the hare IV. American Journal of Botany

101, 1987-2004.

Shimada, H., Sugiura, M., 1991. Fine structural features of the chloroplast genome: comparison of the sequenced chloroplast genomes. Nucleic Acids Res 19, 983-995.

Shinozaki, K., Ohme, M., Tanaka, M., Wakasugi, T., Hayashida, N., Matsubayashi, T., Zaita, N., Chunwongse, J., Obokata, J., Yamaguchi-Shinozaki, K., Ohto, C., Torazawa, K., Meng, B.Y., Sugita, M., Deno, H., Kamogashira, T., Yamada, K., Kusuda, J., Takaiwa, F., Kato, A., Tohdoh, N., Shimada, H., Sugiura, M., 1986. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. The EMBO Journal 5, 2043-2049.

Skog, L.E., Boggan, J.K., 2006. A new classification of the Western Hemisphere Gesneriaceae. Gesneriads 56, 12-19.

Soltis, D.E., Kuzoff, R.K., 1995. Discordance between nuclear and chloroplast phylogenies in the *Heuchera* group (Saxifragaceae). Evolution 49, 727-742.

Soltis, D.E., Soltis, P.S., 1998. Choosing an approach and an appropriate gene for phylogenetic analysis. In: Soltis, D.E., Soltis, P.S., Doyle, J.J. (Eds.), Molecular systematics of plants II. Springer, New York, USA, pp. 1-42.

Song, Y., Yu, W.-B., Tan, Y., Liu, B., Yao, X., Jin, J., Padmanaba, M., Yang, J.-B., Corlett, R.T., 2017. Evolutionary comparisons of the chloroplast genome in Lauraceae and insights into loss events in the Magnoliids. Genome Biol Evol 9, 2354-2364.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312-1313.

Steele, P.R., Hertweck, K.L., Mayfield, D., McKain, M.R., Leebens-Mack, J., Pires, J.C., 2012. Quality and quantity of data recovered from massively parallel sequencing: Examples in Asparagales and Poaceae. American Journal of Botany 99, 330-348.

Straub, S.C.K., Parks, M., Weitemier, K., Fishbein, M., Cronn, R.C., Liston, A., 2012. Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. American Journal of Botany 99, 349-364.

Tanaka, M., Obokata, J., Chunwongse, J., Shinozaki, K., Sugiura, M., 1987. Rapid splicing and stepwise processing of a transcript from the *psbB* operon in tobacco chloroplasts: Determination of the intron sites in *petB* and *petD*. Molecular and General Genetics MGG 209, 427-431.

Testolin, R., Cipriani, G., 1997. Paternal inheritance of chloroplast DNA and maternal inheritance of mitochondrial DNA in the genus *Actinidia*. TAG Theoretical and Applied Genetics 94, 897-903.

Thomson, A.M., Vargas, O.M., Dick, C.W., 2017. Comparative analysis of 24 chloroplast genomes yields highly informative genetic markers for the Brazil nut family (Lecythidaceae). bioRxiv.

Thyssen, G., Svab, Z., Maliga, P., 2012. Exceptional inheritance of plastids via pollen in *Nicotiana sylvestris* with no detectable paternal mitochondrial DNA in the progeny. The Plant Journal 72, 84-88.

Tian, J., Liu, L., Xiao, S.-Y., Zhu, X.-W., Wu, Y.-J., Zhou, L., Yin, J., Pen, L., Cai, X.-Z., Liu, L.-H., Liu, Y.-D., Liu, K.-M., 2018. *Primulina hengshanensis* (Gesneriaceae), a new species from Danxia landform in Hunan, China. 2018 333, 5.

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E.S., Fischer, A., Bock, R., Greiner, S., 2017. GeSeq – versatile and accurate annotation of organelle genomes. Nucleic Acids Res.

Tonti-Filippini, J., Nevill, P.G., Dixon, K., Small, I., 2017. What can we do with 1000 plastid genomes? The Plant Journal 90, 808-818.

Tsitrone, A., Kirkpatrick, M., Levin, D.A., Morgan, M., 2003. A model for chloroplast capture. Evolution 57, 1776-1782.

Turner, B., Munzinger, J., Duangjai, S., Temsch, E.M., Stockenhuber, R., Barfuss, M.H.J., Chase, M.W., Samuel, R., 2013. Molecular phylogenetics of New Caledonian *Diospyros* (Ebenaceae) using plastid and nuclear markers. Molecular Phylogenetics and Evolution 69, 740-763.

Valcárcel, V., Fiz, O., Vargas, P., 2003. Chloroplast and nuclear evidence for multiple origins of polyploids and diploids of *Hedera* (Araliaceae) in the Mediterranean basin.

Molecular Phylogenetics and Evolution 27, 1-20.

Vicient, C.M., Suoniemi, A., Anamthawat-Jónsson, K., Tanskanen, J., Beharav, A., Nevo, E., Schulman, A.H., 1999. Retrotransposon *BARE-1* and Its Role in Genome Evolution in the Genus *Hordeum*. The Plant Cell 11, 1769-1784.

Wang, Y.-Z., Mao, R.-B., Liu, Y., Li, J.M., Dong, Y., Li, Z.-Y., Smith, J.-F., 2011. Phylogenetic reconstruction of *Chirita* and allies (Gesneriaceae) with taxonomic treatments. Journal of Systematics and Evolution 49, 50-64.

Weber, A., 2004. Gesneriaceae. In: Kadereit, J.W. (Ed.), Flowering Plants · Dicotyledons: Lamiales (except Acanthaceae including Avicenniaceae). Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 63-158.

Weber, A., Middleton, D.J., Forrest, A., Kiew, R., Lim, C.-L., Rafidah, A.R., Sontag, S., Triboun, P., Wei, Y.-G., Yao, T.L., Möller, M., 2011a. Molecular systematics and remodelling of *Chirita* and associated genera (Gesneriaceae). Taxon 60, 767-790.

Weber, A., Wei, Y.G., Puglisi, C., Wen, F., Mayer, V., Möller, M., 2011b. A new definition of the genus *Petrocodon* (Gesneriaceae). Phytotaxa 23, 49-67.

Wei, Y.-G., Zhong, S.-H., Wen, H.-Q., 2004. Studies of the flora and ecology Gesneriaceae in Guangxi province. Acta Botanical Yunnanica 26, 173-182.

Wen, F., Liang, G.-Y., Wei, Y.-G., 2012. *Petrocodon lancifolius* (Gesneriaceae), a new species endemic to a central subtropical zone of Guizhou Province, China. 2012 49, 5.

Weng, M.-L., Blazier, J.C., Govindu, M., Jansen, R.K., 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, 645-659.

Wicke, S., Schneeweiss, G.M., 2015. Next-generation organellar genomics: Potentials and pitfalls of high-throughput technologies for molecular evolutionary studies and plant systematics. In: Hörandl, E., Appelhans, M.S. (Eds.), Next-Generation Sequencing in Plant Systematics. Koeltz Scientific Books, Vienna, p. Chapter 1.

Wicke, S., Schneeweiss, G.M., dePamphilis, C.W., Müller, K.F., Quandt, D., 2011. The

evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Molecular Biology 76, 273-297.

Wolf, P.G., 2012. Plastid genome diversity. In: Wendel, J., Greilhuber, J., Dolezel, J., Leitch, I.J. (Eds.), Plant Genome Diversity Volume 1. Springer, pp. 145-154.

Wolfe, K.H., Li, W.H., Sharp, P.M., 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proceedings of the National Academy of Sciences 84, 9054-9058.

Wu, H.-T., Zhou, S.-B., Ma, W., Wang, O.-W., Wen, F., 2017. *Primulina dichroantha* (Gesneriaceae), a New Species from a Karst Cave from Guangxi, China. Annales Botanici Fennici 54, 95-98.

Wyman, S.K., Jansen, R.K., Boore, J.L., 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20, 3252-3255.

Xu, W.-B., Liu, Y., Kono, Y., Chang, H., Peng, C.-I., Chung, K.-F., 2013. *Primulina cardaminifolia* (Gesneriaceae), a rare new species from limestone areas in Guangxi, China. Botanical Studies 24, 19.

Xu, W.-B., Meng, T., Zhang, Q., Wu, W.-H., Liu, Y., Chung, K.-F., 2014. *Petrocodon* (Gesneriaceae) in the limestone karsts of Guangxi, China: three new species and a new combination based on morphological and molecular evidence. Systematic Botany 39, 965-974.

Xu, W.-B., Pan, B., Liu, Y., Peng, C.-I., Chung, K.-F., 2012. Two new species, *Primulina multifida* and *P. pseudomollifolia* (Gesneriaceae), from karst caves in Guangxi, China. Botanical Studies 53, 165-175.

Xu, Z.-R., 1995. A study of the vegetation and floristic affinity of the limestone forests in southern and southwestern China. Ann Mo Bot Gard 82, 570-580.

Yang, L.-H., Pan, B., Kang, M., 2017. *Primulina curvituba* sp. nov. (Gesneriaceae) from a limestone area in Guangxi, China. Nordic Journal of Botany 35, 578-581.

Yang, L.H., Ning, Z.L., Kang, M., 2016. *Primulina rubella* sp. nov. (Gesneriaceae) from a limestone area in Guangdong, China. Nordic Journal of Botany, n/a-n/a.

Yi, D.-K., Kim, K.-J., 2012. Complete Chloroplast Genome Sequences of Important

Oilseed Crop *Sesamum indicum* L. PLOS ONE 7, e35872.

Zerbino, D.R., Birney, E., 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome research 18, 821-829.

Zhang, N., Wen, J., Zimmer, E.A., 2015. Congruent Deep Relationships in the Grape Family (Vitaceae) Based on Sequences of Chloroplast Genomes and Mitochondrial Genes via Genome Skimming. PLOS ONE 10, e0144701.

Zhang, Q., Liu, Y., Sodmergen, 2003. Examination of the Cytoplasmic DNA in Male Reproductive Cells to Determine the Potential for Cytoplasmic Inheritance in 295 Angiosperm Species. Plant and Cell Physiology 44, 941-951.

Zhang, T., Fang, Y., Wang, X., Deng, X., Zhang, X., Hu, S., Yu, J., 2012. The Complete Chloroplast and Mitochondrial Genome Sequences of Boea hygrometrica: Insights into the Evolution of Plant Organellar Genomes. PLOS ONE 7, e30531.

Zhang, X.L., Yang, L.H., Kang, M., 2017. Post-pollination reproductive isolation of sympatric populations of *Primulina eburnea* and *P. mabaensis* (Gesneriaceae). Biodiversity Science 25, 615-620.

Zhu, A., Guo, W., Gupta, S., Fan, W., Mower, J.P., 2016. Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. New Phytologist 209, 1747-1756.
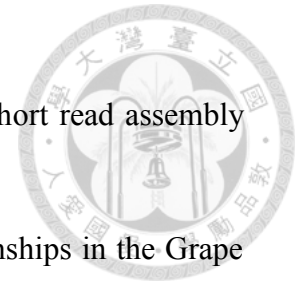
7. Appendix

Table S1. Mitochondrial genes and exons applied in this study.

Gene (exons)	Direction*	Used**
atp1	+	
atp4	-	+
atp6	+	
atp8	-	+
atp9	-	+
ccmB	-	+
ccmC	-	+
ccmFc exon 1	-	+
ccmFc exon 2 exon 3	-	+
ccmFn	+	+
cob	-	+
cox1 exon 1	+	+
cox1 exon 2	+	+
cox2 exon 1	-	+
cox2 exon 2	-	+
cox3	-	+
matR	-	+
nad1 exon 1	-	+
nad1 exon 2	-	+
nad1 exon 3	-	+
nad1 exon 4	+	+
nad2 exon 1	-	+
nad2 exon 2	-	+
nad2 exon 3	-	+
nad2 exon 4	-	+
nad2 exon 5	-	+
nad3	+	+
nad4 exon 1	+	+
nad4 exon 2	+	+
nad4 exon 3	+	+
nad4 exon 4	+	+

*The direction of each exons were marked as "+" and "-" for positive strand and negative strand, respectively.

**The successfully recovered exon used in the phylogenetic was marked as '+'.

Table S1. Mitochondrial genes and exons applied in this study. (continue)

Gene (exons)	Direction*	Used**
nad4L	-	+
nad5 exon 1	+	+
nad5 exon 2	+	+
nad5 exon 3	+	
nad5 exon 4	+	+
nad5 exon 5	+	+
nad6	+	+
nad7 exon 1	+	+
nad7 exon 2	+	+
nad7 exon 3	+	+
nad7 exon 4	+	+
nad9	-	+
rps3 exon 1	-	+
rps3 exon 2	-	+
rps4	-	+
rps7	+	
rps12	+	+
rps13	+	+
sdh3 exon 1	+	+
sdh3 exon 2	+	+

*The direction of each exon was marked as "+" and "-" for positive strand and negative strand, respectively.

**The successfully recovered exon used in the phylogenetic was marked as '+'.

84