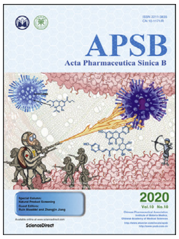
Acta Pharmaceutica Sinica B 2020;10(10):1989–2001

Chinese Pharmaceutical Association
Institute of Materia Medica, Chinese Academy of Medical Sciences

Acta Pharmaceutica Sinica B

www.elsevier.com/locate/apsb
www.sciencedirect.com



ORIGINAL ARTICLE

The large single-copy (LSC) region functions as a highly effective and efficient molecular marker for accurate authentication of medicinal *Dendrobium* species

Ludan Li, Yu Jiang, Yuanyuan Liu, Zhitao Niu, Qingyun Xue, Wei Liu, Xiaoyu Ding*

College of Life Sciences, Nanjing Normal University, Nanjing 210023, China

Received 29 October 2019; received in revised form 19 December 2019; accepted 13 January 2020

KEY WORDS

Dendrobium species of "Fengdou";
Authentication;
Large single-copy (LSC);
Plastomic comparison

Abstract Having great medicinal values, *Dendrobium* species of "Fengdou" (DSFs) are a taxonomically complex group in *Dendrobium* genus including many closely related and recently diverged species. Traditionally used DNA markers have been proved to be insufficient in authenticating many species of this group. Here, we investigated 101 complete plastomes from 23 DSFs, comprising 72 newly sequenced and 29 documented, which all exhibited well-conserved genomic organization and gene order. Plastome-wide comparison showed the co-occurrence of single nucleotide polymorphisms (SNPs) and insertions/deletions (indels), which can be explained by both the repeat-associated and indel-associated mutation hypotheses. Moreover, guanine-cytosine (GC) content was found to be negatively correlated with the three divergence variables (SNPs, indels and repeats), indicating that GC content may reflect the level of the local sequence divergence. Our species authentication analyses revealed that the relaxed filtering strategies of sequence alignment had no negative impact on species identification. By assessing the maximum likelihood (ML) trees inferred from different datasets, we found that the complete plastome and large single-copy (LSC) datasets both successfully identified all 23 DSFs with the maximum bootstrap values. However, owing to the high efficiency of LSC in species identification, we recommend using LSC for accurate authentication of DSFs.

*Corresponding author. Tel./fax: +86 25 85891605.
E-mail address: dingxynj@263.net (Xiaoyu Ding).

Peer review under the responsibility of Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences.

https://doi.org/10.1016/j.apsb.2020.01.012

1. Introduction

Accurate identification of medicinal plants is the basis for their biodiversity conservation and safe utilization. However, effectively identifying species is not an easy task. Morphological features are often under selective pressure[1]. It may result in phenotypic convergence when species adapt to similar growth conditions, while it may also lead to diversification of morphological characters among related species when they adapt to different habitats[1,2]. Traditional DNA barcoding approach does not always track species boundaries due to low interspecific variations of commonly used DNA markers[3–5]. One potential solution to this problem is to increase the amount of data. Recently, it has been suggested that genome-scale datasets should be used to identify the species for which smaller datasets with one or a few DNA regions have limited resolution[6–8]. The nuclear genome contains numerous informative loci for species identification, but obtaining nuclear genome-scale data remains difficult in cost and annotation[9]. By contrast, the plastome has a relatively small size and a high copy number per cell, making whole-plastome sequencing much more feasible[9,10].

The plastomes of nearly all land plants exhibit a typical quadripartite structure with two identical copies of the inverted repeat (IR) separated by a large single-copy (LSC) region and a small single-copy (SSC) region[11]. Owing to their usually uniparental inheritance, moderate evolutionary rate and lack of recombination, plastomes can provide valuable information for taxonomy, species identification and phylogenetic inference[12–14]. In recent years, the rapid development of high-throughput sequencing technologies offers cheaper and simpler access to plastomes than ever before. Consequently, plastomes have been extensively employed to greatly improve phylogenetic resolution and level of species discrimination, particularly in taxonomically complex plant groups, such as Podophylloideae[15], Rosaceae[16], *Echinacea*[17] and *Stipa*[18]. The LSC region, the longest subset of the plastome, can also provide abundant informative sites for phylogenetic analyses and species identification[19,20]. In addition, the highly variable regions of plastomes, which are usually aligned inaccurately, are thought to possibly have a negative impact on phylogenetic inference; hence, their removal might improve robustness of phylogenetic inference[21,22]. However, it has not been reported yet whether ambiguously aligned regions affect species identification.

A case study of the identification of a group of medicinal *Dendrobium* species (*Dendrobium* species of “Fengdou”, DSFs) is presented here. DSFs are an important group in the genus of *Dendrobium* with soft and mucilaginous stems, after being dried and softened, which can be processed into “Fengdou” products. It comprises approximately 20 species in China[23,24]. Being rich in polysaccharides and dendrobine, DSFs have excellent medicinal functions, such as nourishing “Yin”, benefiting the stomach, reducing blood sugar levels and resisting cancer[25]. In the market, many expensive rare DSFs have often been adulterated with other *Dendrobium* species due to their similar appearance. However, the effect of pharmaceutical components greatly differs among *Dendrobium* species[26], so their accurate authentication is vital for medicinal purpose. Accordingly, DNA barcoding has been conducted for DSFs and related species using a single or multi-locus combination sequences[27–30], which, however, are shown to be ineffective in discriminating some important DSFs. For example, the use of the two-marker combination of ITS and *matK* successfully identified most *Dendrobium* species, but failed to distinguish among *D. moniliforme*, *D. fanjingshanense*, *D. officinale*, *D. gratiosissimum* and *D. wardianum*[27]. Recently, in the neighbor-joining tree of ITS2 sequence, *D. huoshanense* and *D. moniliforme* were found to be nested with each other[29]. Moreover, mitochondrial *nad* 1 intron 2 sequences were also utilized to identify nine *Dendrobium* species, and yet could not identify *D. loddigesii*[30]. Therefore, it is urgent to develop an effective and reliable molecular method for authenticating DSFs.

In this study, a total of 101 plastomes from 23 DSFs were analyzed, comprising 72 newly sequenced and 29 previously published. The main aims of this study were: (1) to characterize the plastomes of DSFs regarding genome structure, sequence divergence and guanine-cytosine (GC) content; (2) to assess the potential impact of different filtering strategies of sequence alignment on species identification; (3) to determine a highly effective and efficient molecular method for authentication of DSFs based on the abundant plastomic resources.

2. Materials and methods

2.1. *Taxon sampling and DNA extraction*

We sampled 72 individual plants representing 20 DSFs with 2–10 individuals per species from their main distribution areas (Table 1 and Supporting Information Table S1). All samples were identified by Prof. Xiaoyu Ding, and then grown in the greenhouse of Nanjing Normal University, Nanjing, China. Total genomic DNA of each sample was isolated from fresh leaves (about 500 mg) using the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany) following the manufacturer’s protocol. The quality and concentration of the DNA samples were determined using a DeNovix DS-11 Spectrophotometer (DeNovix Inc, Wilmington, DE, USA). The samples of total genomic DNA with concentration >20 ng/μL, $A_{260}/A_{280} = 1.8–2.0$, and $A_{260}/A_{230}>1.7$ were used for sequencing.

2.2. *DNA sequencing, plastome assembly, annotation and validation*

The total genomic DNA of each tested sample was sequenced using Illumina Hiseq4000 platform (Illumina Inc, San Diego, CA, USA). Approximately 5.0 Gb of raw data was generated with 150 bp paired-end reads for each sample. The raw sequencing reads were trimmed with error probability <0.05, and filtered paired-end reads were assembled on CLC Genomics Workbench v8.5.1 (CLC Bio, Aarhus, Denmark; http://www.clcbio.com) *via* combination of *de novo* and reference-guided assembly approaches following the procedure described by Niu et al.[31] The plastome of *D. officinale* (NC_024019)[32] served as a reference plastome. To validate the assembly, the four junction regions between the IRs and the LSC/SSC were verified by PCR-based conventional Sanger sequencing using specific primers. The finished genomes were annotated using the online program DOGMA v1.2[33]; the start/stop codons and exon/intron boundaries of genes were manually corrected by comparison with homologous genes in the reference genome of *D. officinale*. All tRNA genes were further confirmed using tRNAscan-SE v1.21[34] with default settings. In addition, for each DNA sample, the internal transcribed spacer (ITS) region of nuclear ribosomal DNA was also amplified and sequenced using universal primers provided by Ding et al.[23]

Table 1 Summary of major characteristics of 72 plastomes from 20 *Dendrobium* species of "Fengdou" (DSFs).

No.	Species	Plastome length (bp)	LSC length (bp)	IR length (bp)	SSC length (bp)	GC content (%)				Voucher number	Accession number
						total	LSC	IR	SSC		
1	*D. huoshanense*	151,246	84,868	26,187	14,004	37.53	35.11	43.40	30.30	LLD01_006	LC490207
2	*D. huoshanense*	151,261	84,877	26,187	14,010	37.53	35.11	43.40	30.28	LLD01_052	LC490373
3	*D. huoshanense*	151,281	84,888	26,194	14,005	37.52	35.11	43.36	30.27	LLD01_173	LC490384
4	*D. huoshanense*	151,230	84,898	26,194	13,944	37.52	35.10	43.36	30.33	LLD01_085	LC490375
5	*D. huoshanense*	151,272	84,899	26,194	13,985	37.52	35.10	43.37	30.30	LLD01_139	LC490382
6	*D. huoshanense*	151,279	84,902	26,186	14,005	37.53	35.10	43.39	30.29	LLD01_054	LC490374
7	*D. huoshanense*	151,245	84,881	26,180	14,004	37.53	35.11	43.39	30.28	LLD01_129	LC490380
8	*D. huoshanense*	151,230	84,879	26,180	13,991	37.53	35.11	43.40	30.31	LLD01_114	LC490379
9	*D. huoshanense*	151,269	84,890	26,186	14,007	37.53	35.11	43.39	30.28	LLD01_088	LC490378
10	*D. huoshanense*	151,247	84,878	26,186	13,997	37.53	35.11	43.39	30.31	LLD01_086	LC490381
11	*D. wilsonii*	151,581	84,891	26,236	14,218	37.52	35.09	43.39	30.35	LLD05_023	LC490388
12	*D. wilsonii*	151,590	84,901	26,236	14,217	37.51	35.08	43.39	30.37	LLD05_097	LC490389
13	*D. wilsonii*	151,597	84,891	26,236	14,234	37.51	35.08	43.39	30.31	LLD05_168	LC490392
14	*D. wilsonii*	151,550	84,890	26,236	14,188	37.52	35.08	43.39	30.38	LLD05_231	LC490394
15	*D. wilsonii*	151,579	84,899	26,236	14,208	37.52	35.08	43.39	30.36	LLD05_102	LC490391
16	*D. moniliforme*	150,770	84,875	26,010	13,875	37.54	35.09	43.38	30.58	LLD04_016	LC490386
17	*D. moniliforme*	150,765	84,871	26,010	13,874	37.53	35.08	43.39	30.56	LLD04_153	LC490377
18	*D. moniliforme*	150,755	84,864	26,016	13,859	37.54	35.09	43.38	30.60	LLD04_042	LC490652
19	*D. moniliforme*	150,786	84,878	26,015	13,878	37.53	35.09	43.38	30.53	LLD04_201	LC490387
20	*D. xichouense*	150,772	84,868	26,014	13,876	37.51	35.09	43.38	30.35	LLD09_055	LC490656
21	*D. xichouense*	150,758	84,863	26,008	13,879	37.51	35.09	43.38	30.33	LLD09_192	LC490658
22	*D. xichouense*	150,752	84,839	26,014	13,885	37.51	35.10	43.37	30.31	LLD09_080	LC490657
23	*D. xichouense*	150,745	84,841	26,010	13,884	37.51	35.09	43.36	30.34	LLD09_036	LC490655
24	*D. fanjingshanense*	150,815	84,896	26,035	13,849	37.50	35.08	43.37	30.33	LLD07_106	LC490372
25	*D. fanjingshanense*	150,812	84,885	26,030	13,867	37.52	35.09	43.38	30.35	LLD07_008	LC490405
26	*D. fanjingshanense*	150,827	84,900	26,031	13,865	37.51	35.08	43.38	30.35	LLD07_067	LC490660
27	*D. fanjingshanense*	150,829	84,906	26,029	13,865	37.51	35.08	43.38	30.35	LLD07_069	LC490407
28	*D. fanjingshanense*	150,799	84,887	26,023	13,866	37.51	35.08	43.38	30.35	LLD07_132	LC490653
29	*D. fanjingshanense*	150,772	84,849	26,026	13,871	37.51	35.10	43.37	30.31	LLD07_189	LC490654
30	*D. lituiflorum*	151,232	84,986	26,312	13,622	37.56	35.13	43.38	30.30	LLD23_205	LC490679
31	*D. lituiflorum*	151,225	84,982	26,311	13,621	37.57	35.14	43.38	30.31	LLD23_213	LC490681
32	*D. lituiflorum*	151,230	84,976	26,314	13,626	37.57	35.13	43.38	30.30	LLD23_215	LC490680
33	*D. devonianum*	152,163	85,062	26,289	14,523	37.50	35.08	43.37	30.37	LLD10_064	LC490383
34	*D. devonianum*	152,159	85,053	26,290	14,526	37.50	35.08	43.37	30.39	LLD10_078	LC490385
35	*D. hercoglossum*	152,136	84,988	26,336	14,476	37.49	35.08	43.35	30.33	LLD15_116	LC490400
36	*D. hercoglossum*	152,152	84,962	26,343	14,504	37.53	35.12	43.35	30.50	LLD15_099	LC490398
37	*D. hercoglossum*	152,196	84,994	26,336	14,530	37.48	35.08	43.35	30.23	LLD15_122	LC490402
38	*D. gratiosissimum*	151,803	84,908	26,308	14,279	37.54	35.12	43.37	30.48	LLD17_177	LC490662
39	*D. gratiosissimum*	151,797	84,904	26,308	14,277	37.54	35.12	43.37	30.44	LLD17_156	LC490659
40	*D. gratiosissimum*	151,776	84,898	26,308	14,262	37.55	35.12	43.37	30.49	LLD17_020	LC490406
41	*D. primulinum*	152,931	84,880	26,868	14,315	37.47	35.07	43.23	30.14	LLD08_079	LC490397
42	*D. primulinum*	152,865	84,887	26,848	14,282	37.47	35.08	43.21	30.14	LLD08_072	LC490376
43	*D. primulinum*	152,903	84,917	26,852	14,282	37.46	35.07	43.21	30.13	LLD08_046	LC490399
44	*D. crystallinum*	152,835	84,983	26,890	14,072	37.52	35.11	43.19	30.41	LLD18_134	LC490677
45	*D. crystallinum*	152,865	85,015	26,890	14,070	37.51	35.10	43.19	30.38	LLD18_145	LC490678
46	*D. crystallinum*	152,879	85,027	26,890	14,072	37.51	35.09	43.19	30.42	LLD18_225	LC490676
47	*D. loddigesii*	152,384	84,756	27,027	13,574	37.48	35.06	43.17	29.87	LLD06_061	LC490673
48	*D. loddigesii*	152,368	84,746	27,026	13,570	37.47	35.06	43.17	29.87	LLD06_038	LC490674
49	*D. loddigesii*	152,395	84,760	27,028	13,579	37.46	35.05	43.17	29.82	LLD06_015	LC490396
50	*D. loddigesii*	152,369	84,738	27,028	13,575	37.48	35.07	43.17	29.84	LLD06_029	LC490401
51	*D. aphyllum*	152,462	84,837	27,040	13,545	37.54	35.13	43.20	30.07	LLD13_148	LC490671
52	*D. aphyllum*	152,487	84,857	27,040	13,550	37.55	35.13	43.20	30.10	LLD13_182	LC490669
53	*D. aphyllum*	152,484	84,858	27,037	13,552	37.54	35.11	43.21	30.11	LLD13_186	LC490390
54	*D. falconeri*	153,115	84,987	27,052	14,024	37.44	35.02	43.11	30.26	LLD14_221	LC490408
55	*D. falconeri*	153,124	84,971	27,059	14,035	37.45	35.03	43.11	30.27	LLD14_150	LC490393
56	*D. falconeri*	153,107	84,959	27,059	14,030	37.45	35.04	43.11	30.26	LLD14_208	LC490395
57	*D. wardianum*	153,618	84,997	27,047	14,527	37.48	35.08	43.17	30.29	LLD20_195	LC490661
58	*D. wardianum*	153,625	84,997	27,048	14,532	37.48	35.08	43.17	30.28	LLD20_199	LC490664
59	*D. wardianum*	153,627	84,994	27,048	14,537	37.48	35.08	43.18	30.28	LLD20_210	LC490666
60	*D. lohohense*	153,202	84,932	27,036	14,198	37.49	35.10	43.14	30.35	LLD22_237	LC490670
61	*D. lohohense*	153,175	84,899	27,037	14,202	37.48	35.09	43.13	30.28	LLD22_243	LC490668
62	*D. crepidatum*	153,451	84,933	27,030	14,458	37.49	35.10	43.14	30.38	LLD16_228	LC490675

(*continued on next page*)

Table 1 (*continued*)

No.	Species	Plastome length (bp)	LSC length (bp)	IR length (bp)	SSC length (bp)	GC content (%)				Voucher number	Accession number
						total	LSC	IR	SSC		
63	*D. crepidatum*	153,425	84,915	27,027	14,456	37.48	35.09	43.14	30.37	LLD16_119	LC490403
64	*D. crepidatum*	153,434	84,920	27,029	14,456	37.47	35.08	43.14	30.36	LLD16_130	LC490404
65	*D. chrysanthum*	153,030	84,945	27,030	14,025	37.49	35.07	43.17	30.32	LLD25_222	LC490684
66	*D. chrysanthum*	153,017	84,933	27,030	14,024	37.49	35.08	43.16	30.32	LLD25_256	LC490682
67	*D. chrysanthum*	153,038	84,955	27,030	14,023	37.49	35.06	43.17	30.34	LLD25_260	LC490683
68	*D. pendulum*	152,822	85,006	27,036	13,744	37.53	35.12	43.18	30.44	LLD26_232	LC490663
69	*D. pendulum*	152,787	84,994	27,028	13,737	37.53	35.08	43.18	30.41	LLD26_247	LC490665
70	*D. pendulum*	152,801	85,008	27,028	13,737	37.53	35.09	43.18	30.40	LLD26_267	LC490698
71	*D. strongylanthum*	152,869	84,926	27,022	13,899	37.69	35.25	43.22	31.07	LLD29_281	LC490685
72	*D. strongylanthum*	152,888	84,918	27,022	13,926	37.69	35.27	43.22	31.06	LLD29_272	LC490672

2.3. *Sequence alignment and measurement of divergence variables*

Including 29 documented plastomes of DSFs[31,32,35–38], all of the 101 complete plastome sequences were aligned using the MAFFT v7[39] program under standard parameters. The aligned sequences were partitioned into nonoverlapping bins of 600 bp each. The bins with any sequence completely or mostly missing were removed. Single nucleotide polymorphisms (SNPs), insertions/deletions (indels) and the GC content of each bin were calculated by DnaSP v5.1[40]. Considering the highly conserved feature of plastomes, only one sample of each species was used in the repeat sequence analysis. Forward and reverse repeats with a minimum repeat size of 19 bp and a maximum of one nucleotide mismatch between the two repeat copies in 23 DSFs plastomes were identified using REPuter[41]. The repeats from 23 DSFs plastomes were relocated to each bin according to their locations in plastomes, and repeats shared among species were relocated only once. The number of repeat sequence from each bin was counted. Subsequently, SPSS Statistics 22.0 was employed to determine the correlations between SNPs and indels, SNPs and repeats, indels and repeats, SNPs and GC content, indels and GC content, and repeats and GC content. Moreover, the correlation between SNPs and indels in the bins of coding regions was also assessed through removing the bins with complete noncoding sequences or both coding and noncoding sequences.

2.4. *Extraction of indel-flanking sequences*

Indel-flanking sequences in the alignment of the complete plastomes of DSFs were extracted for examining the distribution of SNPs around indels according to the method of McDonald et al.[42] with some minor modifications. Briefly, 300 bp upstream and downstream sequences of indels were extracted and examined for additional indels. Once a flanking sequence with additional indels was identified, it was immediately removed. Subsequently, a 150 bp sequence adjacent to indels of each flanking sequence was divided into five nonoverlapping bins of 30 bp in size, and then the number of SNP in each bin was calculated. Likewise, indel-flanking sequences in coding regions (excluding the flanking sequences with complete noncoding sequences or both coding and noncoding sequences) were extracted to analyse the relationship between indels and SNPs around indels in coding regions.

2.5. *Filtering strategies of alignments*

Plastomes have many highly variable regions, which are generally aligned with ambiguity. In order to assess the effect of alignment quality on species identification, the datasets of the complete plastome, LSC, IR and SSC were generated. These four datasets of 101 individuals of DSFs and the six outgroup species were aligned using MAFFT v7[39]. Subsequently, the following three filtering strategies were applied to each of the four datasets: the no filtering strategy that retained all sites in the alignment; the light filtering strategy that removed ambiguously aligned regions by using Gblocks v.0.91b[43] with the default parameters, and set allowed-gap-positions at with-half; and the strict filtering strategy that was the same as light filtering strategy except setting allowed-gap-positions at none.

2.6. *Species authentication analyses*

A total of 12 alignments resulted from application of the above-mentioned three filtering strategies to each of the four datasets, which were subsequently used for species authentication. Maximum likelihood (ML) trees of the 12 alignments were reconstructed in RAxML v.8.0.2[44] based on the GTRGAMMA model as suggested (RAxML manual). Bootstrap (BS) values were determined through running 1000 replicates. In addition, to examine whether the four plastome-scale datasets (the complete plastome, LSC, IR and SSC) have higher discriminatory power for DSFs than traditionally used DNA markers, 12 commonly used DNA markers (ITS, ITS2, *matK*, *rbcL*, *trnH-psbA*, *atpF-atpH*, *psbK-psbI*, *trnT-trnL*, *rpl32-trnL*, *clpP-psbB*, *trnL* intron and *rps16-trnQ*) were also used in our authentication study. They were aligned by MUSCLE in MEGA 5.2[45], and all positions containing gaps were removed. Subsequently, they were classified into the following datasets: (1) ITS[23], (2) ITS2[46], (3) ITS + *matK* + *rbcL*[47], (4) ITS2+*matK* + *rbcL*[47], (5) ITS2+*trnH-psbA*[48], (6) *matK* + *rbcL*[49], (7) *matK* + *trnH-psbA*[50], (8) *rbcL* + *trnH-psbA*[51], (9) *matK* + *trnH-psbA* + *atpF-atpH*[52], (10) *matK* + *atpF-atpH* + *psbK-psbI*[52], (11) *trnH-psbA* + *atpF-atpH* + *psbK-psbI*[53], and (12) *trnT-trnL* + *rpl32-trnL* + *clpP-psbB* + *trnL* intron + *rps16-trnQ*[35]. The combinations of DNA markers were concatenated in SequenceMatrix v1.7.8[54]. Likewise, these datasets were used for the ML tree analyses. If all

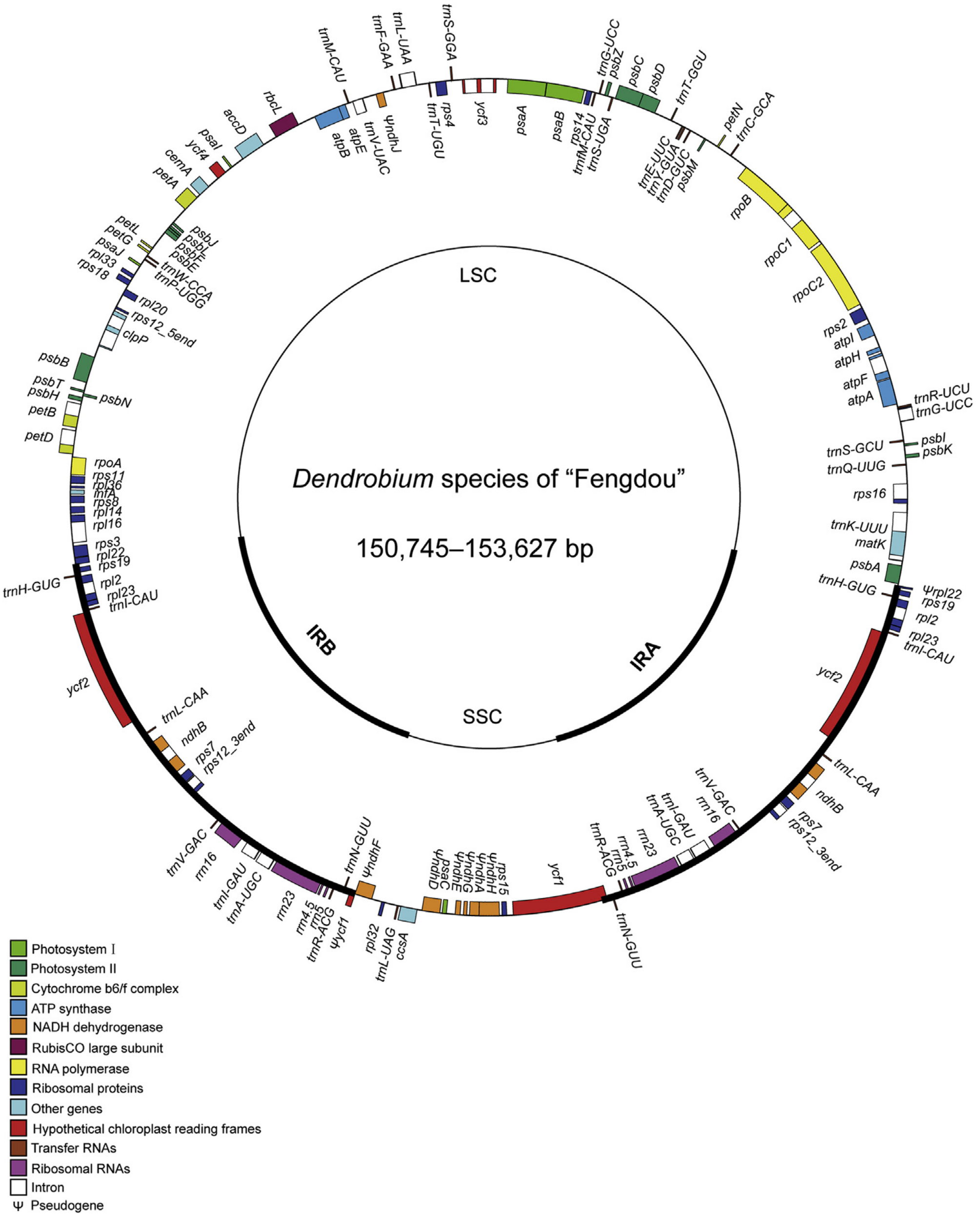


Figure 1 Plastome map of *Dendrobium* species of “Fengdou”. The genes inside and outside the circle are transcribed clockwise and counterclockwise, respectively. Genes from different functional groups are shown in different colors. The thick lines represent the inverted repeat regions (IR_A and IR_B) that separate the plastome into large single-copy (LSC) and small single-copy (SSC) regions.

individuals of one species were clustered into a monophyletic clade with a bootstrap value above 70%, then the species was considered to be successfully identified[5].

3. Results

3.1. General features of new plastomes

The 72 newly sequenced plastomes of DSFs ranged in length from 150,745 (*D. xichouense*) to 153,627 bp (*D. wardianum*, Table 1). All these plastomes displayed a typical quadripartite structure consisting of a pair of IR regions (26,008−27,059 bp) separated by the LSC (84,738−85,062 bp) and SSC (13,545−14,537 bp) regions. The overall GC contents in 72 DSFs plastomes were 37.44%−37.69%, whereas those in the LSC, IR, and SSC regions were 35.02%−35.27%, 43.11%−43.40% and 29.82%−31.07%, respectively. These plastomes each consistently contained 103 unique genes, which were arranged in the same order across the plastomes (Fig. 1). The 103 unique genes consisted of 69 protein-coding genes (CDS), 30 tRNA genes and four rRNA genes. Of these, eight CDS (*rps6*, *atpF*, *rpoC1*, *petB*, *petD*, *rpl16*, *rpl2* and *ndhB*) each contained a single intron, as did six tRNA genes (*trnK*^{UUU}, *trnG*^{UCC}, *trnL*^{UAA}, *trnV*^{UAC}, *trnI*^{GAU} and *trnA*^{UGC}), while three CDS (*ycf3*, *clpP* and *rps12*) each possessed two introns. Interestingly, the *rps12* gene was *trans*-spliced; the 5′ end exon lay in the LSC region, while the 3′ end exon and intron were located in the IR regions. Besides, nine pseudogenes (*ψndhA*, *D*, *E*, *F*, *G*, *H*, *J*, *ψrpl22* and *ψycf1*) were identified in all DSFs plastomes.

DSFs plastomes were compared for IR/SC boundaries and their adjacent genes (Fig. 2). Although the gene content and gene order were well-conserved across these plastomes, obvious differences at the IR/SC boundaries were still observed. The IR_B region expanded into the *rpl22* gene, leading to the duplication of a 37 bp fragment of *ψrpl22* at the IR_A/LSC border. The *ycf1* gene crossed the IR_A/SSC region, resulting in the duplication of a 91−1076 bp fragment of *ψycf1* in the IR_B region. Furthermore, the *ψndhF* gene was found to vary in size from 256 to 1899 bp, and *ψndhF* and *ψycf1* overlapped by 3−72 bp at the IR_B/SSC border among the DSFs plastomes. These results revealed obvious expansion or contraction of the IRs in the plastomes of DSFs.

3.2. *Correlations among SNPs, indels, repeats and GC content in the plastomes of DSFs*

The three divergence variables (SNPs, indels and repeats) and GC content were assessed in the 101 DSFs plastomes. At whole-plastome level, a total of 260 bins contained 7259 SNPs, 2980 indels and 2432 forward and reverse repeats (Supporting Information Table S2). The percentages of divergence variables and GC content in each bin are visualized in a line plot (Fig. 3). The divergence variables were nonrandomly distributed in different regions of DSFs plastomes. The IR regions showed lower level of variability than the SC regions. Besides, this plot showed close correlations among SNPs, indels, repeats and GC content. Correlation coefficients (*r*) for each pair of the parameters were determined (Table 2). The pairwise comparisons revealed that positive correlations existed between SNPs and indels, between SNPs and repeats, and between indels and repeats. The degree of correlation between indels and repeats was greatest, followed by that between SNPs and indels, which in turn exceeded that between SNPs and repeats. On the other hand, these divergence variables were all negatively correlated with GC content. The strongest correlation occurred between SNPs and GC content, followed by that between repeats and GC content, which in turn was stronger than that between indels and GC content. All the correlation coefficients were significant at $P < 0.01$. These results showed close associations among SNPs, indels and repeats. The distribution of these divergence variables may be dependent on the level of the local GC content.

3.3. *SNP density around indels*

A total of 223 indel-flanking sequences were extracted from the alignment of the complete plastomes of DSFs, 81 of which were located in coding regions. To estimate the effect of indels on the distribution of nearby SNPs, a jackknife resampling approach was applied to randomly extract 120 and 50 flanking sequences from complete plastomes and coding regions, respectively, with 1000 iterations to calculate average SNP density within each bin of 150 bp sequence adjacent to indels. Fig. 4 shows the relationships between SNP density and the distance to indels at the two levels. In both cases, SNP density decreased as the distance to indels increased. The most rapid decline in SNP density occurred in the first few bins closest to indels. These results suggested that the distance to indels had a strong effect on SNP density.

3.4. *Species authentication analyses*

To assess the effect of the different filtering strategies on species identification, average bootstrap values supporting the monophyly of 23 DSFs were calculated. Average bootstrap values and sequence alignment characteristics of the four datasets are presented in Table 3. The unfiltered complete plastome dataset had an aligned length of 179,210 bp with 25,763 variable sites and 12,376 parsimony informative sites, of which 67.2% variable sites and 67.5% parsimony informative sites originated from the unfiltered LSC dataset. In all the datasets except the SSC, the percentages of variable and parsimony informative sites in strictly filtered alignments slightly declined compared with the respective counterparts in either the unfiltered or the lightly filtered alignment.

For the LSC and complete plastome datasets, the unfiltered and lightly filtered alignments both exhibited the greatest discriminatory power of 100% for DSFs, as did the strictly filtered alignments (Table 3, Supporting Information Figs. S1 and S2). In the six ML trees derived from these two datasets, all the individuals of each species were clustered into a monophyletic clade with highest bootstrap value. Differently, the strictly filtered IR and SSC datasets were insufficient for identification of all DSFs; in contrast, both the unfiltered and lightly filtered alignments of the two datasets showed better species resolution with higher average bootstrap values (Table 3, Supporting Information Figs. S3 and S4). On the other side, commonly used DNA markers were found to have limited discriminating power for DSFs (Fig. 5 and Supporting Information Fig. S5). In the 12 datasets of commonly used DNA markers (Materials and methods section), ITS + *matK* + *rbcL* and *trnH-psbA* + *atpF-atpH* + *psbK-psbI* showed the highest (87.0%) and the lowest species discrimination rate (52.2%), respectively, while the core barcode *matK* + *rbcL* recommended by CBOL just exhibited moderate species resolution rate (73.9%). These commonly used DNA markers appeared to be ineffective in identifying the following species: *D. huoshanense*, *D. wilsonii*, *D. moniliforme*, *D. fanjingshanense*, *D. xichouense*, *D. wardianum* and *D. chrysanthum*. They were often nested with other species or formed monophyletic groups with low support values in ML trees inferred from some of these datasets, *e.g.*, ITS, ITS2, *matK* + *rbcL*, *rbcL* + *trnH-psbA* and *matK* + *trnH-psbA*. Altogether, these results indicated that the LSC and complete plastome datasets could effectively distinguish among all of the tested DSFs, and the different filtering strategies (no, light and strict) made no difference to their authentication results.

4. Discussion

4.1. *Mutational dynamics of the plastome*

The co-occurrence of substitutions and indels is generally observed in prokaryote and eukaryote genomes[55,56]. Recently, this phenomenon has also been reported in plant chloroplast genomes[57−59]. Three major hypotheses have been suggested to explain this phenomenon, consisting of the repeat-associated mutation hypothesis[42], the indel-associated mutation hypothesis[55,56] and the regional difference hypothesis[60,61]. However, these hypotheses have not been explicitly investigated in *Dendrobium* plastomes. Here, we examined and discussed the three hypotheses, and explored the mutational dynamics of the plastome based on the data of a large number of *Dendrobium* plastomes.

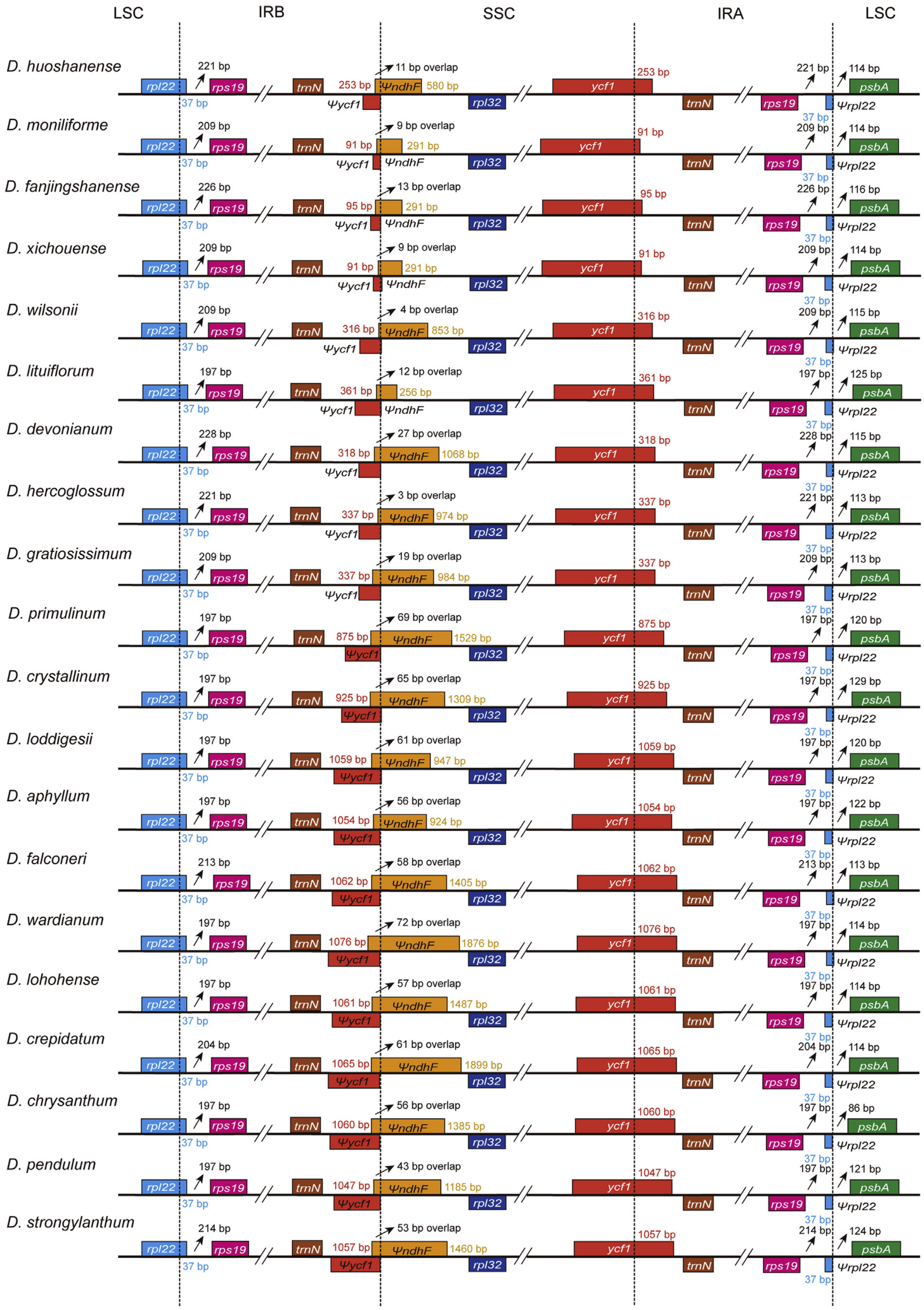


Figure 2 Comparison of IR/SC junction regions among the plastomes of 20 *Dendrobium* species of "Fengdou".

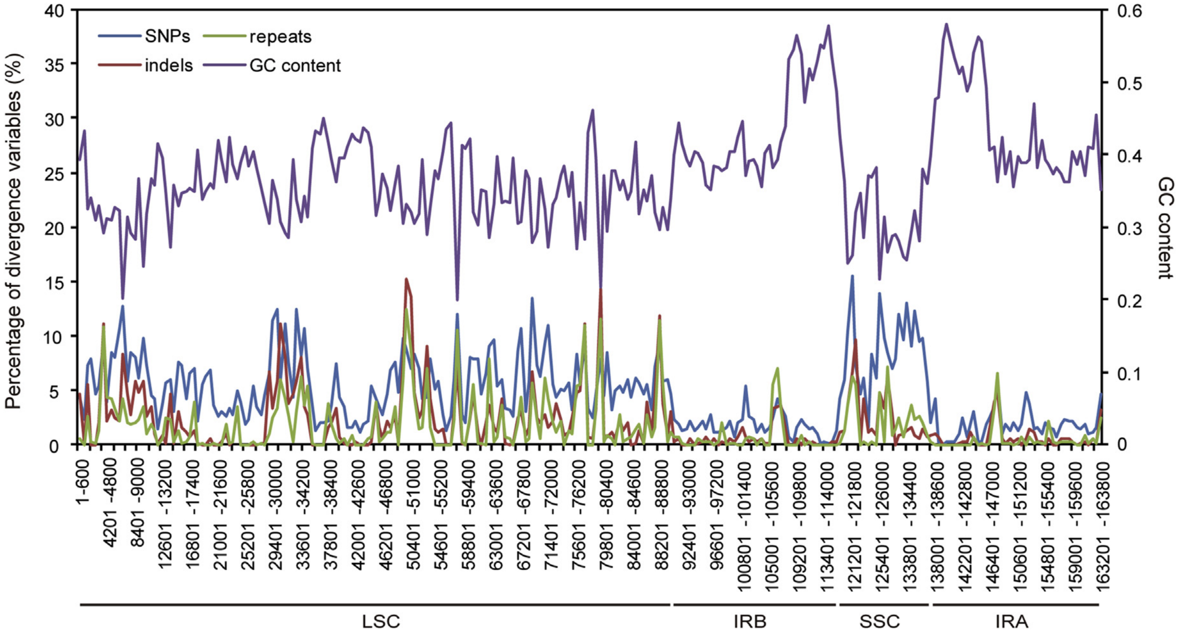


Figure 3 Percentages of SNPs, indels, and repeats and GC content in 260 nonoverlapping bins of 600 bp each through the alignment of 101 complete plastomes of 23 *Dendrobium* species of "Fengdou".

Table 2 Correlations between each pair of the parameters SNPs, indels, repeats and GC content in the alignment of 101 complete plastomes of 23 DSFs.

Comparison	Correlation coefficient (*r*)	*P*-value
SNPs and indels	0.705	**
SNPs and repeats	0.688	**
Indels and repeats	0.750	**
SNPs and GC content	−0.814	**
Indels and GC content	−0.640	**
Repeats and GC content	−0.690	**

The alignment was partitioned into 260 nonoverlapping bins of 600 bp each to calculate correlation coefficients. **Stands for that correlation was significant at $P < 0.01$ level (two-tailed).

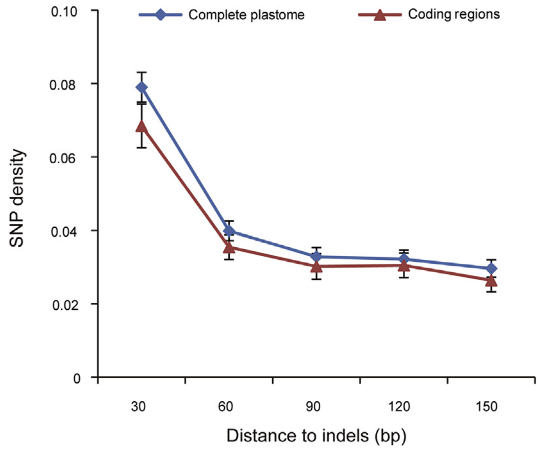


Figure 4 Relationships between SNP density and the distance to indels at the levels of the complete plastome and coding regions.

4.1.1. *The repeat-associated mutation hypothesis*

The repeat-associated mutation hypothesis[42] states that repeats tend to increase the mutation rates of both nucleotide substitution and indel in surrounding sequences, which has been supported by many investigators. For example, Ahmed et al.[57,58] and Yi et al.[59] reported the existence of the genome-wide associations among repeats, indels and substitutions in both Aroid and *Cephalotaxus* plastomes, demonstrating that repeats play an important role in inducing substitution and indel mutations. Consistently, this study showed strong associations among SNPs, indels and repeats, which confirmed the co-occurrence of SNPs and indels in *Dendrobium* plastomes, and also provided new evidence to further support the repeat-associated mutation hypothesis[42]. Mechanistically, repeats are prone to induce nucleotide substitutions through recruiting the error-prone DNA polymerases during DNA replication[42]; besides, they also tend to increase the likelihood of indels by slipped-strand mispairing[62,63]. Thus, repeats act as a common cause for indels and substitutions, which is a potential explanation for the association between substitutions and indels.

4.1.2. *The indel-associated mutation hypothesis*

In addition to repeats, indels *per se* may also function as a mutator to induce SNPs, directly resulting in their association. The present study showed that SNP density was higher at positions closer to indels, indicating a strong impact of indels on the nearby SNP density. A similar distribution pattern of nucleotide diversity around indels was also reported in prokaryotes and eukaryotes[55,56]. These consistent observations pointed to the indel-associated mutation hypothesis[55,56]. The potential mechanism for this hypothesis is that heterozygous indels are expected to enhance local replication errors, causing the occurrence of nucleotide mutations at nearby sites[56,64]. In other words, the assumed mutagenic effect of indels is dependent on their heterozygosity, indicating that heterozygous indels possibly play a more important role in molecular and genome evolution than homozygous ones.

Table 3 Sequence alignment characteristics and average bootstrap values supporting the monophyly of 23 DSFs under different alignment filtering strategies.

Dataset	Alignment	Number of sites	Variable sites (%)	Parsimony informative sites (%)	Average bootstrap value (%)
Complete plastome	Unfiltered	179,210	25,763 (14.4)	12,376 (6.9)	100
	Lightly filtered	148,077	20,585 (13.9)	10,424 (7.0)	100
	Strictly filtered	120,787	13,363 (11.1)	6465 (5.4)	100
LSC	Unfiltered	99,126	17,313 (17.5)	8356 (8.4)	100
	Lightly filtered	82,830	14,272 (17.2)	7197 (8.7)	100
	Strictly filtered	67,153	9220 (13.7)	4497 (6.7)	100
IR	Unfiltered	28,213	1686 (6.0)	723 (2.6)	94.0
	Lightly filtered	26,004	1503 (5.8)	629 (2.4)	95.4
	Strictly filtered	22,317	1050 (4.7)	442 (2.0)	86.5
SSC	Unfiltered	22,637	5347 (23.6)	2706 (12.0)	99.3
	Lightly filtered	12,362	3493 (28.3)	1983 (16.0)	99.7
	Strictly filtered	7760	1973 (25.4)	1014 (13.1)	97.0

When calculating average bootstrap values, the bootstrap values for non-monophyletic species were considered to be zero.

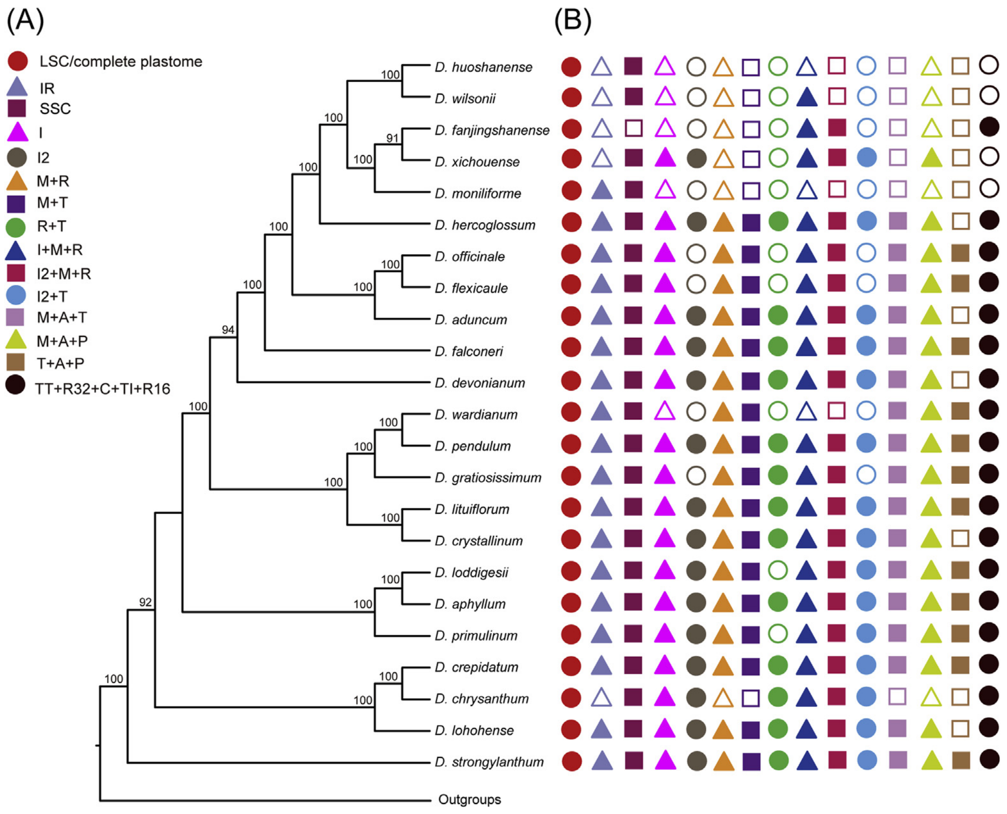


Figure 5 Comparisons of the discriminatory power of 16 datasets for 23 *Dendrobium* species of "Fengdou". (A) Maximum likelihood (ML) tree of 23 *Dendrobium* species of "Fengdou" inferred from the strictly filtered alignment of LSC dataset. All individuals of each species formed a monophyletic clade, but only one individual per species is shown for simplicity. Bootstrap support values >50% are shown above branches. (B) Identification results of different datasets for 23 *Dendrobium* species of "Fengdou" (Figs. S1−S5). The identification results of the complete plastome, LSC, IR and SSC were derived from their strictly filtered datasets. Different datasets are shown in different colors and shapes. For each dataset, the solid figures represent successfully identified species, whereas the hollow ones stand for unsuccessfully identified species (I, ITS; I2, ITS2; M, *matK*; R, *rbcL*; T, *trnH-psbA*; A, *atpF-atpH*; P, *psbK-psbI*; TT, *trnT-trnL*; R32, *rpl32-trnL*; C, *clpP-psbB*; TI, *trnL* intron; R16, *rps16-trnQ*).

4.1.3. *The regional difference hypothesis*

It is well known that the mutations in coding regions are probably more deleterious than those in noncoding regions. Under purifying selection, mutation events such as SNPs and indels may be disproportionately distributed in noncoding regions[55]; thus, the regional difference may cause the observed association between indels and SNPs at the plastomic level. To test this possibility, we investigated the correlation between SNPs and indels in coding regions, which was shown to be significant (Spearman's $r = 0.415$, $P < 0.01$), though weaker than that at the complete plastome level. Moreover, the distribution pattern of SNPs around indels in coding regions was similar to that in the complete plastome (Fig. 4). These results suggested that the regional difference hypothesis cannot sufficiently explain the association between SNPs and indels, hence leaving the first two hypotheses as the potential explanations for the observed association.

4.1.4. *A high level of sequence divergence associated with low GC content*

In addition to testing the three hypotheses mentioned above, the current study also examined the relationship between GC content and the three divergence variables SNPs, indels and repeats. Interestingly, these variables were all negatively correlated with GC content. This raised the possibility that these divergence variables vary together only because they all change with the fluctuations in GC content. Recently, a similar correlation between GC content and the extent of sequence variation was also observed in Apostasioideae plastomes[65]. These findings indicated that the mutation events in Orchid plastomes may be dependent on GC-poor composition. GC-poor hyper-variable regions were also detected in rice[66] and cycad[67]. Therefore, GC content may become an indicator of the levels of the local divergence variables; in turn, the levels of divergence variables may also provide information about the GC content. Altogether, our study demonstrated the existence of the complex interactions among SNPs, indels, repeats and GC content in plastomes; however, the mechanisms driving their interactions remain to be elucidated for deeply understanding plastome evolution.

4.2. *The influence of alignment filtering on species authentication*

The tree-based methods have become particularly popular in species identification because of their sensitivity, robustness and intuitiveness[68,69]. As new sequencing technologies continue to bring about an exponential increase in DNA sequence output, species identification and phylogenetic studies have entered the era of genome. In genome-scale datasets, inaccurate alignments caused by hyper-variable regions are considered to potentially mislead phylogenetic inference[16,21,22], yet it had not been evaluated whether the quality of sequence alignment can affect species identification based on the tree-building approaches. This study investigated the impact of alignment quality on species authentication by applying the three filtering strategies (no, light and strict) to each of the four datasets (the complete plastome, LSC, IR and SSC). For both the LSC and complete plastome datasets, each of the three filtering strategies successfully identified all the tested DSFs with maximum support values. On the other side, for both the IR and SSC datasets, their strictly filtered alignments showed a lower discriminatory power for DSFs than the unfiltered and lightly filtered alignments, which was mainly due to that many useful informative sites were removed by the strict filtering strategy[70]. So according to our results, neither the unfiltered strategy nor the lightly filtered one had a negative impact on species identification. In fact, relaxed filtering strategies brought better results for shorter datasets. Recently, the influence of alignment quality on phylogeny has been assessed in Cornales[71]. Similar to our results, different filtering strategies were shown to make no difference to phylogenetic resolution, which was mainly attributed to the conservation of plastomes within Cornales. Our findings are instrumental in identifying appropriate filtering strategies of sequence alignment for species identification. When the adopted alignment is long enough and contains sufficient informative sites, it does not matter which filtering strategy is selected. By contrast, if the alignment used is relatively short, in order to avoid losing many useful informative sites, a relaxed filtering strategy is more appropriate.

4.3. *The LSC region is recommended for accurate authentication of DSFs*

In *Dendrobium*, DSFs are a taxonomically complex group characterized by many closely related and recently diverged species[28,72,73]. It is notoriously difficult to authenticate them. Traditional methods for discriminating among DSFs are based on their morphological characters, while overlapping interspecific variations lead to inadequate diagnostic characteristics for their identification[38,74]. Furthermore, previous molecular identification studies based on one or a few DNA regions were also proved to be ineffective in authenticating many important DSFs[27,29,30]. Consistent with previous results, in this study, single markers or multi-marker combinations were demonstrated to be unable to identify all of the tested DSFs. Recent phylogenetic studies have shown that mainland Asian *Dendrobium* is a recent radiation group[28,72]. The limited species resolution of the commonly used DNA markers may be due to the lack of accumulated variations among evolutionarily young groups[75]. The complete plastome sequences contain massive informative sites, which have been shown to be more effective in identifying taxonomically difficult taxa[17,18,76,77]. More recently, Zhu et al.[36] successfully distinguished among *D. officinale* and its closely related species using the complete plastome sequences, which showed great advantages of the complete plastome sequences in discriminating among closely related species. However, the efficacy of this approach in authenticating other taxonomically complex taxa of *Dendrobium* remains to be assessed. Here, we employed the complete plastome sequences to authenticate DSFs based on a large sampling scale of 23 species, with multiple individuals sampled for each species; as a result, they effectively differentiated among all the tested DSFs. Therefore, our results further confirmed the effectiveness of the complete plastome sequences in identifying *Dendrobium* species.

Among the subsets of the plastome, the LSC region is the largest in size, accounting for more than a half of the entire length of the plastome. Owing to the advantage of its length, the LSC region contains a majority of the informative sites of the complete plastome (about 67%, Table 3), therefore having the potential for becoming an alternative to the complete plastome. Indeed, many studies have recently demonstrated that the LSC region has almost the same performance as the complete plastome in species identification and phylogenetic studies[14,19,20]. Consistently, in our study, the LSC region exhibited the highest discrimination power for DSFs, as did the complete plastome, implying that the LSC region, containing main variable sites for identification of DSFs, can substitute for the complete plastome to discriminate among DSFs. On the other hand, we also screened the 6 relatively highly variable fragments of LSC region for DSFs (*rps16-trnQ*, *ndhJ-trnV*, *atpB-rbcL*, *psbB-psbT*, *trnT-psbD* and *trnT-trnL*, Supporting Information Fig. S6), and tested whether the combination of multiple taxon-specific markers can effectively identify DSFs. Unfortunately, this combination was still insufficient in authenticating all the tested DSFs (Supporting Information Fig. S7), further suggesting that the extended DNA barcode approach should be used to distinguish among DSFs.

In fact, compared to the complete plastome, the LSC region has many advantages in species identification. Firstly, its relatively short matrix from samples is easier to align and requires less storage space and computational time, especially when involving huge sample sets. Secondly, if only for the purpose of species identification, we can directly map the trimmed reads to the LSC region of the reference plastome and merely assemble the LSC regions of samples, which mean that plastome annotation and PCR verification of IR/SC junction regions can be omitted. Thirdly, because of its stable structure, the LSC region is relatively easy to assemble and requires less data depth, which can reduce sequencing cost to some extent. Overall, the use of the LSC region for species identification is more time-saving, effort-saving and cost-effective than the utilization of the complete plastome. Therefore, the LSC region is recommended for accurate authentication of DSFs.

5. Conclusions

This work is the first discrimination among DSFs, a taxonomically complex group, based on a large-scale plastome sampling. Firstly, a comparison across 101 DSFs plastomes revealed a close association between SNPs and indels, which can be explained by both the repeat-associated and the indel-associated mutation hypotheses. Furthermore, all the three divergence variables (SNPs, indels and repeats) were found to be negatively correlated with GC content, implying that GC content may serve as an indicator of the levels of these divergence variables. Most importantly, species authentication analyses using the ML tree-building method demonstrated that the no or light filtering strategy did not adversely affect the authentication results. While the LSC and the complete plastome datasets both showed the highest discriminatory power of 100% for DSFs, the LSC has many potential advantages over the complete plastome in the efficiency and cost of authentication. Therefore, we recommend using the LSC for rapid and accurate identification of DSFs.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 31170300 and 31670330), and the Projects Funded by Jiangsu Agricultural Independent Innovation (Grant No. CX(18)3063, China) and the Priority Academic Program Development of Jiangsu Higher Education Institutions, Nanjing, China.

Author contributions

Xiaoyu Ding designed the study. Ludan Li, Yu Jiang, Yuanyuan Liu and Wei Liu performed the experiments. Ludan Li, Yu Jiang and Yuanyuan Liu analysed the data. Ludan Li wrote the manuscript. Xiaoyu Ding, Zhitao Niu and Qingyun Xue revised the manuscript. All of the authors approved the final version of the manuscript.

Conflicts of interest

The authors declare no conflict of interest.

Appendix A. Supporting information

Supporting data to this article can be found online at https://doi.org/10.1016/j.apsb.2020.01.012.

References

1. Koopman WJM, Wissemann V, de Cock K, van Huylenbroeck J, de Riek J, Sabatino GJ, et al. AFLP markers as a tool to reconstruct complex relationships: a case study in *Rosa* (Rosaceae). *Am J Bot* 2008;**95**:353−66.
2. Welker CAD, Souza-Chies TT, Longhi-Wagner HM, Peichoto MC, McKain MR, Kellogg EA. Multilocus phylogeny and phylogenomics of *Eriochrysis* P. Beauv. (Poaceae−Andropogoneae): taxonomic implications and evidence of interspecific hybridization. *Mol Phylogenet Evol* 2016;**99**:155−67.
3. Chen J, Zhao JT, Erickson DL, Xia N, Kress WJ. Testing DNA barcodes in closely related species of *Curcuma* (Zingiberaceae) from Myanmar and China. *Mol Ecol Resour* 2015;**15**:337−48.
4. Wang XM, Gussarova G, Ruhsam M, de Vere N, Metherell C, Hollingsworth PM, et al. DNA barcoding a taxonomically complex hemiparasitic genus reveals deep divergence between ploidy levels but lack of species-level resolution. *AoB Plants* 2018;**10**:ply026.
5. Li YL, Tong Y, Xing FW. DNA barcoding evaluation and its taxonomic implications in the recently evolved genus *Oberonia* Lindl. (Orchidaceae) in China. *Front Plant Sci* 2016;**7**:1791.
6. Li XW, Yang Y, Henry RJ, Rossetto M, Wang YT, Chen SL. Plant DNA barcoding: from gene to genome. *Biol Rev* 2015;**90**:157−66.
7. Coissac E, Hollingsworth PM, Lavergne S, Taberlet P. From barcodes to genomes: extending the concept of DNA barcoding. *Mol Ecol* 2016; **25**:1423−8.
8. Hollingsworth PM, Li DZ, van der Bank M, Twyford AD. Telling plant species apart with DNA: from barcodes to genomes. *Philos Trans R Soc Lond B Biol Sci* 2016;**371**:20150338.
9. Yan MH, Fritsch PW, Moore MJ, Feng T, Meng AP, Yang J, et al. Plastid phylogenomics resolves infrafamilial relationships of the Styracaceae and sheds light on the backbone relationships of the ericales. *Mol Phylogenet Evol* 2018;**121**:198−211.
10. Tonti-Filippini J, Nevill PG, Dixon K, Small I. What can we do with 1000 plastid genomes?. *Plant J* 2017;**90**:808−18.
11. Zhu AD, Guo WH, Gupta S, Fan WS, Mower JP. Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. *New Phytol* 2016;**209**: 1747−56.
12. Zhang YJ, Du LW, Liu A, Chen JJ, Wu L, Hu WM, et al. The complete chloroplast genome sequences of five *Epimedium* species: lights into phylogenetic and taxonomic analyses. *Front Plant Sci* 2016;**7**:306.
13. Xin TY, Zhang Y, Pu XD, Gao RR, Xu ZC, Song JY. Trends in herbgenomics. *Sci China Life Sci* 2019;**62**:288−308.
14. Yang JB, Tang M, Li HT, Zhang ZR, Li DZ. Complete chloroplast genome of the genus *Cymbidium*: lights into the species identification, phylogenetic implications and population genetic analyses. *BMC Evol Biol* 2013;**13**:84.
15. Ye WQ, Yap ZY, Li P, Comes HP, Qiu YX. Plastome organization, genome-based phylogeny and evolution of plastid genes in Podophylloideae (Berberidaceae). *Mol Phylogenet Evol* 2018;**127**:978−87.
16. Zhang SD, Jin JJ, Chen SY, Chase MW, Soltis DE, Li HT, et al. Diversification of Rosaceae since the late cretaceous based on plastid phylogenomics. *New Phytol* 2017;**214**:1355−67.
17. Zhang N, Erickson DL, Ramachandran P, Ottesen AR, Timme RE, Funk VA, et al. An analysis of *Echinacea* chloroplast genomes: implications for future botanical identification. *Sci Rep* 2017;**7**:216.
18. Krawczyk K, Nobis M, Myszczyński K, Klichowska E, Sawicki J. Plastid super-barcodes as a tool for species discrimination in feather grasses (Poaceae: *Stipa*). *Sci Rep* 2018;**8**:1924.

19. Du YP, Bi Y, Yang FP, Zhang MF, Chen XQ, Xue J, et al. Complete chloroplast genome sequences of *Lilium*: insights into evolutionary dynamics and phylogenetic analyses. *Sci Rep* 2017;**7**:5751.
20. Dong M, Zhou XM, Ku WZ, Xu ZG. Detecting useful genetic markers and reconstructing the phylogeny of an important medicinal resource plant, *Artemisia selengensis*, based on chloroplast genomics. *PLoS One* 2019;**14**:e0211340.
21. Zhong BJ, Deusch O, Goremykin VV, Penny D, Biggs PJ, Atherton RA, et al. Systematic error in seed plant phylogenomics. *Genome Biol Evol* 2011;**3**:1340−8.
22. Som A. Causes, consequences and solutions of phylogenetic incongruence. *Brief Bioinf* 2015;**16**:536−48.
23. Ding XY, Wang ZT, Xu H, Xu LS, Zhou KY. Database establishment of the whole rDNA ITS region of *Dendrobium* species of "Fengdou" and authentication by analysis of their sequences. *Acta Pharm Sin* 2002;**37**:567−73.
24. Geng LX, Zheng R, Ren J, Niu ZT, Sun YL, Xue QY, et al. Application of new type combined fragments: nrDNA ITS+*nad* 1-intron 2 for identification of *Dendrobium* species of Fengdous. *Acta Pharm Sin* 2015;**50**:1060−7.
25. Chinese Pharmacopoeia Committee. *Chinese pharmacopoeia*. Beijing: China Medical Science and Technology Press; 2015.
26. Lam Y, Ng TB, Yao RM, Shi J, Xu K, Sze SC, et al. Evaluation of chemical constituents and important mechanism of pharmacological biology in *Dendrobium* plants. *Evid Base Compl Alternat Med* 2015; **2015**:841752.
27. Xu SZ, Li DZ, Li JW, Xiang XG, Jin WT, Huang WC, et al. Evaluation of the DNA barcodes in *Dendrobium* (Orchidaceae) from mainland Asia. *PLoS One* 2015;**10**:e0115168.
28. Xiang XG, Schuiteman A, Li DZ, Huang WC, Chung SW, Li JW, et al. Molecular systematics of *Dendrobium* (Orchidaceae, Dendrobieae) from mainland Asia based on plastid and nuclear sequences. *Mol Phylogenet Evol* 2013;**69**:950−60.
29. Wang XY, Chen XC, Yang P, Wang LL, Han JP. Barcoding the *Dendrobium* (Orchidaceae) species and analysis of the intragenomic variation based on the internal transcribed spacer 2. *BioMed Res Int* 2017;**2017**:2734960.
30. Zhang T, Wang ZT, Xu LS, Zhou KY. Application of mitochondrial *nad* 1 intron 2 sequences to molecular identification of some species of *Dendrobium* Sw. *Chin Tradit Herb Drugs* 2005;**36**:1059−62.
31. Niu ZT, Xue QY, Wang H, Xie XZ, Zhu SY, Liu W, et al. Mutational biases and GC-biased gene conversion affect GC content in the plastomes of *Dendrobium* Genus. *Int J Mol Sci* 2017;**18**:2307.
32. Luo J, Hou BW, Niu ZT, Liu W, Xue QY, Ding XY. Comparative chloroplast genomes of photosynthetic orchids: insights into evolution of the Orchidaceae and development of molecular markers for phylogenetic applications. *PLoS One* 2014;**9**:e99016.
33. Wyman SK, Jansen RK, Boore JL. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* 2004;**20**:3252−5.
34. Schattner P, Brooks AN, Lowe TM. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res* 2005;**33**:686−9.
35. Niu ZT, Zhu SY, Pan JJ, Li LD, Sun J, Ding XY. Comparative analysis of *Dendrobium* plastomes and utility of plastomic mutational hotspots. *Sci Rep* 2017;**7**:2073.
36. Zhu SY, Niu ZT, Xue QY, Wang H, Xie XZ, Ding XY. Accurate authentication of *Dendrobium officinale* and its closely related species by comparative analysis of complete plastomes. *Acta Pharm Sin B* 2018;**8**:969−80.
37. Niu ZT, Xue QY, Zhu SY, Sun J, Liu W, Ding XY. The complete plastome sequences of four orchid species: insights into the evolution of the Orchidaceae and the utility of plastomic mutational hotspots. *Front Plant Sci* 2017;**8**:715.
38. Niu ZT, Pan JJ, Xue QY, Zhu SY, Liu W, Ding XY. Plastome-wide comparison reveals new SNV resources for the authentication of *Dendrobium huoshanense* and its corresponding medicinal slice (Huoshan Fengdou). *Acta Pharm Sin B* 2018;**8**:466−77.
39. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 2013;**30**:772−80.
40. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 2009;**25**:1451−2.
41. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. Reputer: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res* 2001;**29**:4633−42.
42. McDonald MJ, Wang WC, Huang HD, Leu JY. Clusters of nucleotide substitutions and insertion/deletion mutations are associated with repeat sequences. *PLoS Biol* 2011;**9**:e1000622.
43. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 2000;**17**:540−52.
44. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014;**30**:1312−3.
45. Tamura K. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 2011;**28**:2731−9.
46. Chen SL, Yao H, Han JP, Liu C, Song JY, Shi LC, et al. Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PLoS One* 2010;**5**:e8613.
47. Li DZ, Gao LM, Li HT, Wang H, Ge XJ, Liu JQ, et al. Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proc Natl Acad Sci U S A* 2011;**108**:19641−6.
48. Chen SL, Pang XH, Song JY, Shi LC, Yao H, Han JP, et al. A renaissance in herbal medicine identification: from morphology to DNA. *Biotechnol Adv* 2014;**32**:1237−44.
49. Cbol Plant Working Group. A DNA barcode for land plants. *Proc Natl Acad Sci U S A* 2009;**106**:12794−7.
50. Newmaster SG, Fazekas AJ, Steeves RAD, Janovec J. Testing candidate plant barcode regions in the Myristicaceae. *Mol Ecol Resour* 2008;**8**:480−90.
51. Kress WJ, Erickson DL. A two-locus global DNA barcode for land plants: the coding *rbcL* gene complements the non-coding *trnH-psbA* spacer region. *PLoS One* 2007;**2**:e508.
52. Pennisi E. Taxonomy. Wanted: a barcode for plants. *Science* 2007;**318**: 190−1.
53. Kim HM, Oh SH, Bhandari GS, Kim CS, Park CW. DNA barcoding of orchidaceae in Korea. *Mol Ecol Resour* 2014;**14**:499−507.
54. Vaidya G, Lohman DJ, Meier R. SequenceMatrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon information. *Cladistics* 2011;**27**:171−80.
55. Zhu LC, Wang Q, Tang P, Araki H, Tian DC. Genomewide association between insertions/deletions and the nucleotide diversity in bacteria. *Mol Biol Evol* 2009;**26**:2353−61.
56. Tian DC, Wang Q, Zhang PF, Araki H, Yang SH, Kreitman M, et al. Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature* 2008;**455**:105−8.
57. Ahmed I, Biggs PJ, Matthews PJ, Collins LJ, Hendy MD, Lockhart PJ. Mutational dynamics of aroid chloroplast genomes. *Genome Biol Evol* 2012;**4**:1316−23.
58. Ahmed I, Matthews PJ, Biggs PJ, Naeem M, McLenachan PA, Lockhart PJ. Identification of chloroplast genome loci suitable for high-resolution phylogeographic studies of *Colocasia esculenta* (L.) Schott (Araceae) and closely related taxa. *Mol Ecol Resour* 2013;**13**:929−37.
59. Yi X, Gao L, Wang B, Su YJ, Wang T. The complete chloroplast genome sequence of *Cephalotaxus oliveri* (Cephalotaxaceae): evolutionary comparison of *Cephalotaxus* chloroplast DNAs and insights into the loss of inverted repeat copies in gymnosperms. *Genome Biol Evol* 2013;**5**:688−98.
60. Silva JC, Kondrashov AS. Patterns in spontaneous mutation revealed by human-baboon sequence comparison. *Trends Genet* 2002;**18**:544−7.
61. Hardison RC, Roskin KM, Yang S, Dieknans M, Kent WJ, Weber R, et al. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res* 2003;**13**:13−26.

62. Levinson G, Gutman GA. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol* 1987;**4**:203−21.
63. Kelchner SA. The evolution of non-coding chloroplast DNA and its application in plant systematics. *Ann Mo Bot Gard* 2000;**87**:482−98.
64. Longman-Jacobsen N, Williamson JF, Dawkins RL, Gaudieri S. In polymorphic genomic regions indels cluster with nucleotide polymorphism: quantum Genomics. *Gene* 2003;**312**:257−61.
65. Niu ZT, Pan JJ, Zhu SY, Li LD, Xue QY, Liu W, et al. Comparative analysis of the complete plastomes of *Apostasia wallichii* and *Neuwiedia singapureana* (Apostasioideae) reveals different evolutionary dynamics of IR/SSC boundary among photosynthetic Orchids. *Front Plant Sci* 2017;**8**:1713.
66. Yamane K, Yano K, Kawahara T. Pattern and rate of indel evolution inferred from whole chloroplast intergenic regions in sugarcane, maize and rice. *DNA Res* 2006;**13**:197−204.
67. Wu CS, Chaw SM. Evolutionary stasis in cycad plastomes and the first case of plastome GC-biased gene conversion. *Genome Biol Evol* 2015;**7**:2000−9.
68. Manzanilla V, Kool A, Nguyen Nhat L, Nong Van H, Le Thi Thu H, de Boer HJ. Phylogenomics and barcoding of *Panax*: toward the identification of ginseng species. *BMC Evol Biol* 2018;**18**:44.
69. Liu J, Milne RI, Möller M, Zhu GF, Ye LJ, Luo YH, et al. Integrating a comprehensive DNA barcode reference library with a global map of yews (*Taxus* L.) for forensic identification. *Mol Ecol Resour* 2018;**18**:1115−31.
70. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* 2007;**56**:564−77.
71. Fu CN, Li HT, Milne R, Zhang T, Ma PF, Yang J, et al. Comparative analyses of plastid genomes from fourteen Cornales species: inferences for phylogenetic relationships and genome evolution. *BMC Genom* 2017;**18**:956.
72. Xiang XG, Mi XC, Zhou HL, Li JW, Chung SW, Li DZ, et al. Biogeographical diversification of mainland Asian *Dendrobium* (Orchidaceae) and its implications for the historical dynamics of evergreen broad-leaved forests. *J Biogeogr* 2016;**43**:1310−23.
73. Hou BW, Luo J, Zhang YS, Niu ZT, Xue QY, Ding XY. Iteration expansion and regional evolution: phylogeography of *Dendrobium officinale* and four related taxa in southern China. *Sci Rep* 2017;**7**:43525.
74. Yukawa T, Uehara K. Vegetative diversification and radiation in subtribe Dendrobiinae (Orchidaceae): evidence from chloroplast DNA phylogeny and anatomical characters. *Plant Syst Evol* 1996;**201**:1−14.
75. Ruhsam M, Rai HS, Mathews S, Ross TG, Graham SW, Raubeson LA, et al. Does complete plastid genome sequencing improve species discrimination and phylogenetic resolution in *Araucaria*?. *Mol Ecol Resour* 2015;**15**:1067−78.
76. Chen XL, Zhou JG, Cui YX, Wang Y, Duan BZ, Yao H. Identification of *Ligularia* herbs using the complete chloroplast genome as a super-barcode. *Front Pharmacol* 2018;**9**:695.
77. Hu H, Hu QJ, Al-Shehbaz IA, Luo X, Zeng TT, Guo XY, et al. Species delimitation and interspecific relationships of the genus *Orychophragmus* (Brassicaceae) inferred from whole chloroplast genomes. *Front Plant Sci* 2016;**7**:1826.