Phylogenetic inference of Ericales based on plastid genomes and implication of cp-SSRs

ANJAN HAZRA [1], SUBHANWITA DAS [2], SENJUTI BHATTACHARYA [2],
SUSMITA SUR [1], CHANDAN SENGUPTA [3], SAUREN DAS [1] *

[1] Indian Statistical Institute, Kolkata, West Bengal, India

[2] Department of Molecular Biology and Biotechnology, University of Kalyani, Nadia, India

[3] Department of Botany, University of Kalyani, Nadia, India

Abstract

Ericales is an ancient eudicot order encompassing numerous species of economic and ornamental values. Despite several phylogenomic studies, the evolutionary relationship among certain families of this group remains uncertain. The present study assessed a multilocus species tree of Ericales based on 107 chloroplast genomes. The plastome derived microsatellite motifs were also simultaneously explored to check their dynamicity in corroboration of species phylogeny and systematics. In addition to resolving the usual hierarchy, the present phylogenetic analysis enabled to resolve the persisting lineage disparity with valid statistical support. Accordingly, divergence incongruences of Primulaceae, Ebenaceae, and Sapotaceae from earlier reports were reinstated in presently inferred phylogeny, which further supported the latest transcriptome-based relationship of the corresponding group. Various SSR motif characteristics emerged following the recognition of the evolutionary pathway. Numerical variation in tetranucleotide repeats showed even intraspecific or varietal differences in *Camellia sinensis*. Validation of plastome microsatellite-based polymorphism among the related taxa might pave the way for future phylogenetic and population studies of this economically important group.

Key words: chloroplast genome, Ericales, phylogeny, simple sequence repeats, species tree

Introduction

Ericales is a diverse angiosperm clade with an ancient lineage, and it comprises ~12,000 species distributed in 21 or 22 families (Rose et al., 2018). Most of the plants in this group are woody perennials. Notably, a major population in this group is commercially utilized, including tea (*Camellia sinensis*), kiwifruit (*Actinidia deliciosa*), blueberry, huckleberry, cranberry, etc. Few species such as *Cyclamen*, *Impatiens*, *Polyanthus*, primroses (Primulaceae), and rhododendrons (Ericaceae) are renowned for their fabulous aesthetic value (Larson et al., 2020). Recently, several approaches have been used to recon the relationship among the clades across Ericales through both chloroplast loci (Rose et al., 2018) and transcriptome dataset-based matrices (Larson et al., 2020). However, incongruences in phylogenetic relationships still persist for some clades between both the

analyses, which is probably due to sampling limitations and/or in commutable datasets (Larson et al., 2020). In particular, the rapid evolutionary divergence during several million years has made it difficult to infer the monophyly of several clades within this eudicot order (Anderberg et al., 2002; Rose et al., 2018; Schönenberger et al., 2005).

Molecular marker-assisted phylogenetic and systematic analyses have become a preferential method in plant science since the last two decades (Daniell et al., 2021; Gitzendanner et al., 2018; Hazra et al., 2020; Hazra et al., 2018; Soltis and Soltis, 2021). The presence of potential variable characteristics within the whole chloroplast genomes of plant species enables to infer a better resolution in plant relationships (Dong et al., 2014; Dong et al., 2012; Douglas, 1998; Huang et al., 2014; Zhao et al., 2015). Simple sequence repeats (SSRs) are a type of 1–6 nu-

cleotide unit tandem repeat sequence motifs that are frequently observed in chloroplast genomes (cp-Genomes). Chloroplast SSRs (cp-SSRs) are unevenly distributed in the plastomes across various taxa, and because of their high polymorphism, they have potential applications in phylogenetics and breeding programs (Du et al., 2017; Perdereau et al., 2017; Tong et al., 2016). Hence, they are often considered as effective molecular markers for detecting intraspecific or interspecific polymorphisms (Pauwels et al., 2012; Powell et al., 1995; Provan et al., 1997). Documentation of appropriate biomarkers, together with morphological traits, for Ericales is always required to preferentially support the reliable authentication of species. This helps to accurately identify all closely or distantly related taxa for conservation or economic purposes.

Chloroplast SSR-mediated marker development might simplify the process of identification and track the phylogeny of certain taxa. The Gen Bank database includes 111 reference chloroplast genome sequences belonging to various taxa under the order Ericales (last accessed May 28, 2019). However, according to current knowledge, genome-wide phylogenetic comparisons have not been conducted for spanning this sequenced group. In the present study, phylogenetic inference of Ericales has been drawn through combined plastid loci and the SSR motifs from plastomes of diverse lineages (107 taxa belonging to 12 families) to evaluate their phylogenetic implications.

Methods

Data retrieval

Complete plastid genome nucleotide sequences of all available taxa under Ericales were retrieved from NCBI GenBank (Table S1). A total of 111 reference genomes were found covering 12 families. Seven genomes belonging to the family *Ericaceae* were excluded from further analysis due to incomplete or erratic sequence length. Additionally, the complete chloroplast genomes of *Camellia sinensis* var. *dehungensis*, *C. sinensis* var. *pubilimba*, and *C. sinensis* var. *assamica* were included in the dataset, and plastomes of five taxa under the superasterid order Caryophyllales were included in the analyses as an out-group. Thus, the final dataset for phylogenetic inference consisted of 107 Ericales taxa and five out-group species (Table S1).

Phylogenetic analyses

All protein-coding gene sequences within cp-Genomes of the studied taxa were extracted from the NCBI database. DNA sequences of single-copy protein-coding genes from each species were aligned using the MUSCLE program (Edgar, 2004) with default parameters implemented in Mesquite v3.31 (Maddison, 2008). The best-fit substitution models for phylogenetic inference were selected by running jModel Test 3.7 (Posada, 2008). Maximum likelihood tree analyses were performed using RAxML v.8.2.12 (Stamatakis, 2014) at the CIPRES Science Gateway website (Miller et al., 2010). For ML tree inference, the best-fit model, *i.e.*, general time-reversible (GTR) + G + I, was used. A 1000 round bootstrap analysis was included in the study, and the bootstrap supports are shown on the nodes.

Mining SSR motifs

The SSR loci within the 107 Ericales plastid genomes were identified using the MISA tool (Beier et al., 2017). The minimum numbers of repeats motifs were considered as 10, 6, 4, 3, 3, and 3 for mono-, di-, tri-, tetra-, penta-, and hexanucleotides, respectively. The number of repeats in each taxon was presented along the phylogenetic hierarchy of Ericales and then compared.

Results and discussion

Chloroplast genome resources of Ericales in Gen Bank included 114 unique taxa, with Theaceae being the predominant taxa (Table S1). Following manual filtering, the final dataset of complete genomes contained 107 taxa representing 11 families, namely Theaceae (61), Primulaceae (11), Actinidiaceae (7), Ebenaceae (7), Styracaceae (6), Pentaphylacaceae (5), Symplocaceae (3), Lecythidaceae (3), Balsaminaceae (2), Sapotaceae (1), and Sladeniaceae (1). Mining and comparison of the protein-coding genes among all the sampled taxa generated a set of 55 genes without any missing data. These 55 genes were then selected for concatenated phylogenetic analyses (Table S2). Following multiple sequence alignment of single-copy genes, the range of alignment lengths was between 90 nt (*petN*) and 8468 nt (*ycf2*). Similarly, the highest amount of parsimony informative sites were observed in the alignment of *ycf2* (3181) and the lowest amount was recorded in *petN* (9) (Fig. 1, Table S2). Subsequently, on the basis of the alignment
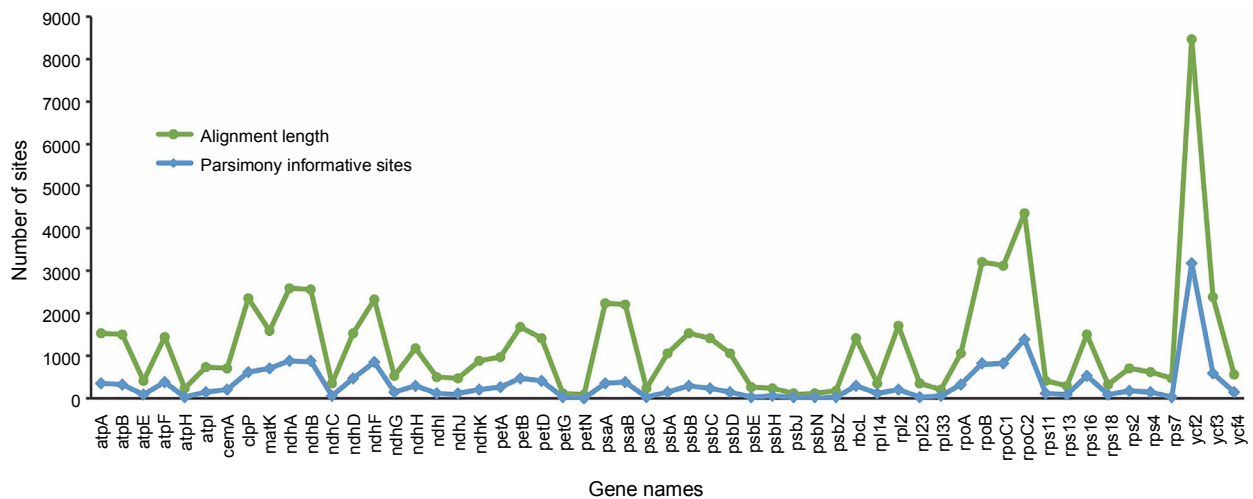
Fig. 1. Comparison of nucleotide variability from the alignment of protein-coding genes considered for phylogenetic analyses

of concatenated 55 genes, the maximum likelihood phylogenetic relationship among the Ericales taxa was elucidated (Fig. 2). The resultant species tree showed monophyly of Theaceae, Actinidiaceae, Symplocaceae, Strycaceae, Pentaphylaceae, Sapotaceae, Ebenaceae, Lecythidaceae, Primulaceae, and Balsaminaceae with significant bootstrap support (>90%). Here, the Balsaminaceae lineage was observed to be more primitive than all other clades, and this corroborates with earlier reports on Ericales phylogeny (Rose et al., 2018). Within Theaceae, all the genera with multiple species representatives, such as *Stewartia, Schima, Polyspora* and *Camellia* were clustered in a single clade. The final topology of the species tree in the present analysis supported Sapotaceae as the sister of the core Ericales group and that Primulaceae has diverged earlier than Lecythidaceae and Ebenaceae (Fig. 2). This observation agrees with the earlier report (Larson et al., 2020), which stated that Sapotaceae and Ebenaceae are sister to each other and they together are sister to the core Ericales. Interestingly, similar to the nuclear loci-based earlier phylogeny (Larson et al., 2020), the findings of the chloroplast genome-assisted analysis of the present study supported earlier divergence event of Primulaceae with strong bootstrap support (=100%). Moreover, Primulaceae does not share the same clade with Sapotaceae and Ebenaceae, unlike suggested earlier by most of the previous phylogenetic studies (Gitzendanner et al., 2018; Rose et al., 2018).

Variable repeat motifs (such as SSRs) in cp-Genomes of angiosperm plants, originating due to slipped-strand mispairing, play a significant role in sequence rearrangement and ultimately cause plastome-wide variations across taxa (Huotari and Korpelainen, 2012; Raubeson et al., 2007; Yuan et al., 2017; Zhang et al., 2016). Accordingly, these repeat polymorphism events would be seemingly useful in plant phylogenetic studies (Cavalier-Smith, 2002; Nie et al., 2012). Here, among all the types of chloroplast microsatellite motifs, mononucleotide repeats (≥ 10 units) were the most predominant form in each taxon (Fig. 2, Table S1). However, the number of such repeat motifs varied among lineages (Table 1, Table S1). Among the species of Theaceae and Actinidiaceae, these repeats are evenly distributed; however, the distribution varied widely among members of families such as Styracaceae, Ebenaceae and Primulaceae (Table 1). Du et al. (2017) also reported that mononucleotide repeats are the most abundant type in the chloroplast genomes of *Lilium* species, and these markers along with other repeat motifs might be potentially useful in population studies. Moreover, the mononucleotide repeats were mostly A/T type (Table S3),which is consistent with earlier similar reports (Kuang et al., 2011; Yin et al., 2018).

In the present study, dinucleotide repeats were unique to some specific lineages and were completely absent from *Camellia, Polyspora*, and *Diospyros* (Table 1). Likewise, penta- and hexanucleotide repeat motifs were taxa specific, the occurrence of which was unique to Actinidiaceae, Ebenaceae and Styracaceae (Table 1, Table S3). Tri- and tetranucleotide repeats were common in the plastomes of most of the studied taxa and
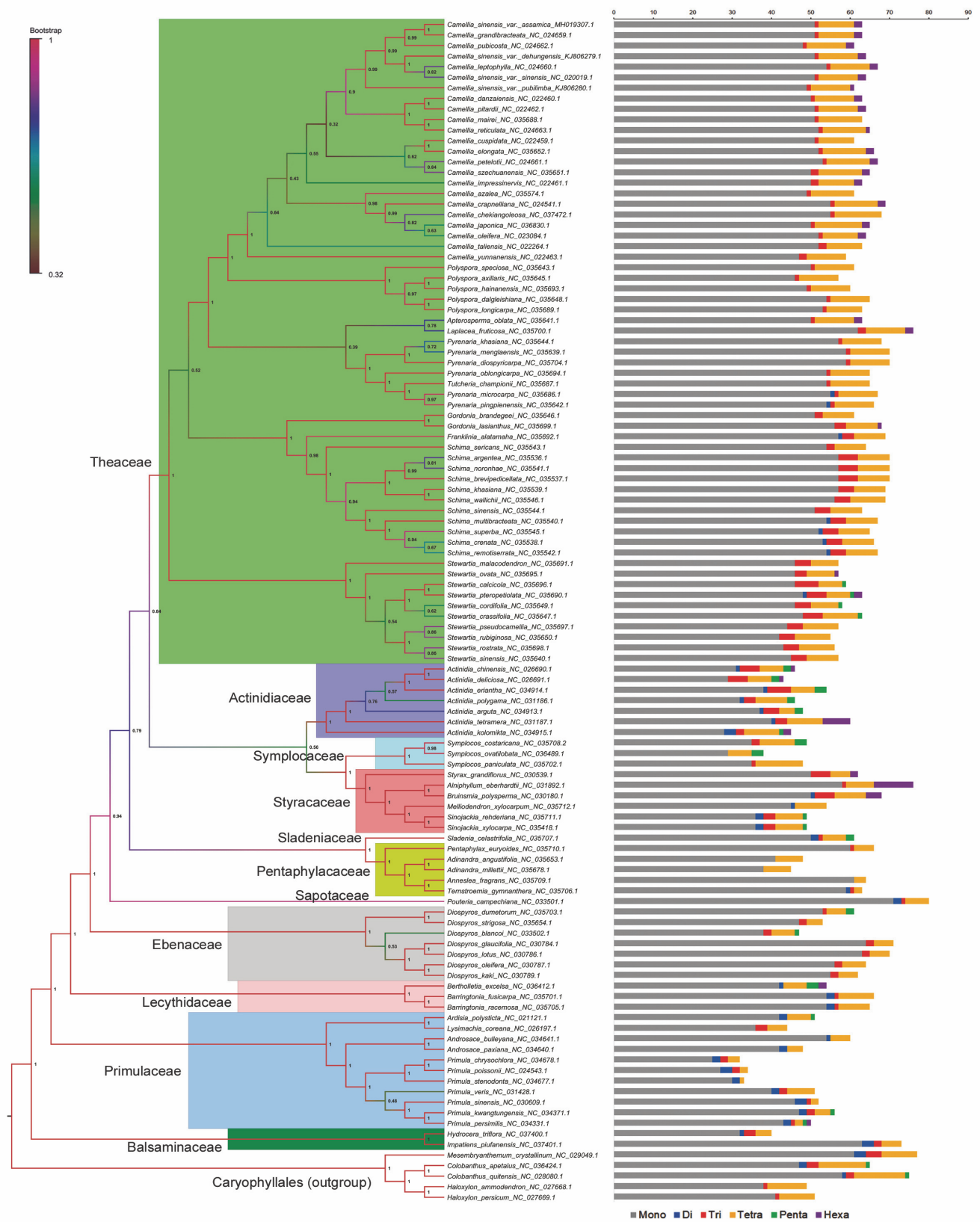
Fig. 2. Phylogenetic relationship of 107 Ericales taxa inferred from maximum likelihood analyses based on the concatenated matrix of 55 protein-coding genes; five Caryophyllales species have been taken as an outgroup; bootstrap supports frequency (0–1) for each node used for the branch color of the tree; statistical results of various repeat motifs extracted from the chloroplast genome of the corresponding taxa have been provided alongside

Table 1. Frequency ranges of various SSR motifs (di-, tri-, tetra, penta- and hexanucleotides)
across plastid genomes of the studied Ericales taxa

Family and genus	Occurrences of various SSR motifs in cp-Genomes				
	Di	Tri	Tetra	Penta	Hexa
Actinidiaceae	0–3	2–6	4–9	1–3	0–7
Actinidia	0–3	2–6	4–9	1–3	0–7
Balsaminaceae	1–3	2–3	4–5	0	0
Hydrocera	1	3	4	0	0
Impatiens	3	2	5	0	0
Ebenaceae	0	1–2	4–6	1–2	0
Diospyros	0	1–2	4–6	1–2	0
Lecythidaceae	1–2	0–1	6–9	0–3	0–2
Barringtonia	2	1	8–9	0	0
Bertholletia	1	0	6	3	2
Pentaphylacaceae	0–1	0–1	2–7	0	0
Adinandra	0	0	7	0	0
Anneslea	0	0	3	0	0
Pentaphylax	0	1	5	0	0
Ternstroemia	1	1	2	0	0
Primulaceae	0–3	1–3	1–7	0–1	0–1
Androsace	1–2	0	4–5	0	0
Ardisia	2	0	6	1	0
Lysimachia	0	3	5	0	0
Primula	2–3	1–2	2–4	0–1	0–1
Sapotaceae	2	1	6	0	0
Pouteria	2	1	6	0	0
Sladeniaceae	2	1	6	2	0
Sladenia	2	1	6	2	0
Styracaceae	0–2	0–5	5–8	0–1	0–10
Alniphyllum	0	1	7	0	10
Bruinsmia	1	5	8	0	4
Melliodendron	1	0	8	0	0
Sinojackia	2	3	7	1	0
Styrax	0	5	5	0	2
Symplocaceae	0	0–2	6–12	0–3	0
Symplocos	0	0–2	6–12	0–3	0
Theaceae	0–1	1–6	6–12	0–1	0–2
Apterosperma	0	1	10	0	2
Camellia	0	1–2	9–12	0	0–2
Franklinia	1	3	8	0	0
Gordonia	0	2–3	8	0	0–1
Laplacea	0	2	10	0	2
Polyspora	0	1	9–10	0	0
Pyrenaria	0–1	1	10	0	0
Schima	0–1	2–5	8–9	0	0
Stewartia	0–1	3–6	6–9	0–1	0–2
Tutcheria	0	1	10	0	0
Varietal difference in *Camellia sinensis*					
C. sinensis var. *assamica*	0	1	9	0	2
C. sinensis var. *sinensis*	0	1	10	0	2
C. sinensis var. *pubilimba*	0	1	10	0	1

Table 2. Signature SSR motifs that differentiate
C. sinensis var. *assamica* and *C. sinensis* var. *sinensis* at the intraspecies level

Taxa	Accession	SSR type	SSR motif	Size	Start position	End position
C. sinensis var. *assamica*	MH019307.1	tetranucleotide	(TCTT)3	12	34 114	34 125
C. sinensis var. *sinensis*	NC_020019.1		(CCCT)3	12	110 118	110 129
			(GAGG)3	12	133 619	133 630

were distributed with varying frequencies. Here, the unique distribution of various types of repeat motifs in the monophyletic groups such as Actinidiaceae and Theaceae (Table 1, Table S1) represents their synapomorphic existence. For example, tetranucleotide repeat motifs could help in differentiating *Camellia* species at their varietal level (Table 1, Table S3). *Camellia sinensis* var. *assamica* plastome possessed 9 tetranucleotide motifs, whereas *C. sinensis* var. *sinensis* had 10 tetranucleotide motifs (Table S1). The unique tetranucleotide repeats differentiating *C. sinensis* and *C. assamica* plastomes are summarized in Table 2. These findings are consistent with the earlier reports of similar analyses which indicated that SSR polymorphisms in plastomes might serve as potential molecular markers for perceiving both intra- and interspecific delimitation (Asaf et al., 2017; Pauwels et al., 2012; Powell et al., 1995; Provan et al., 1997). Subsequently, polymorphic markers can be developed from these repeat flanking regions of organellar genome, which on further validation, may serve as an efficient phylogenetic and systematic tool as exemplified earlier in *Lilium* and *Fritillaria* (Bi et al., 2018; Du et al., 2017).

Conclusion

Plastid genome-based phylogenetic reconstruction of 107 Ericales taxa spanning 12 families showed precise resolution of some phylogenetically conflicted group, especially the relationship among Primulaceae, Ebenaceae and Sapotaceae. The dynamics of chloroplast microsatellites supports phylogenetic relationships among the studied taxonomic groups. Occurrences of tetranucleotide motifs could also differentiate varieties of *Camellia sinensis*. Overall, the findings of this study enabled to clearly establish phylogenetic relationships in Ericales, a diverse angiosperm order, by using plastidial markers.

References

Anderberg A.A., Rydin C., Källersjö M. (2002) *Phylogenetic relationships in the order Ericales sl: analyses of molecular data from five genes from the plastid and mitochondrial genomes*. Amer. J. Bot. 89(4): 677–687.

Asaf S., Khan A.L., Aaqil Khan M., Muhammad Imran Q., Kang S.-M., Al-Hosni K., Jeong E.J., Lee K.E., Lee I.-J. (2017) *Comparative analysis of complete plastid genomes from wild soybean (Glycine soja) and nine other Glycine species*. PloSone 12(8): e0182281.

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

Bi Y., Zhang M.-F., Xue J., Dong R., Du Y.-P., Zhang X.-H. (2018) *Chloroplast genomic resources for phylogeny and DNA barcoding: a case study on Fritillaria*. Sci. Rep. 8(1): 1–12.

Cavalier-Smith T. (2002) *Chloroplast evolution: secondary symbiogenesis and multiple losses*. Curr. Biol. 12(2): R62–R64.

Daniell H., Jin S., Zhu X.G., Gitzendanner M.A., Soltis D.E., Soltis P.S. (2021) *Green giant − a tiny chloroplast genome with mighty power to produce high value proteins: history and phylogeny*. Plant Biotechnol. J. 19(3): 430–447.

Dong W., Liu H., Xu C., Zuo Y., Chen Z., Zhou S. (2014) *A chloroplast genomic strategy for designing taxon specific DNA mini-barcodes: a case study on ginsengs*. BMC Gen. 15(1): 1–8.

Dong W., Liu J., Yu J., Wang L., Zhou S. (2012) *Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding*. PloSone 7(4): e35071.

Douglas S.E. (1998) *Plastid evolution: origins, diversity, trends*. Curr. Opin. Genet. Develop. 8(6): 655–661.

Du Y.-P., Bi Y., Yang F.-P., Zhang M.-F., Chen X.-Q., Xue J., Zhang X.-H. (2017) *Complete chloroplast genome sequences of Lilium: insights into evolutionary dynamics and phylogenetic analyses*. Sci. Rep. 7(1): 1–10.

Gitzendanner M.A., Soltis P.S., Wong G.K.S., Ruhfel B.R., Soltis D.E. (2018) *Plastid phylogenomic analysis of green plants: a billion years of evolutionary history*. Am. J. Bot. 105(3): 291–301.

Hazra A., Bhowmick S., Sengupta C., Das S. (2020) *Lowest copy nuclear genes in disentangling plant molecular systematics*. Taiwania 65(4): 413–422.

Hazra A., Nandy P., Sengupta C., Das S. (2018) *MIPS sequences: a promising molecular consideration in angiosperm phylogeny and systematics*. BioTechnologia 99(1): 5–12.

Huang H., Shi C., Liu Y., Mao S.-Y., Gao L.-Z. (2014) *Thirteen Camelliachloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships*. BMC Evol. Biol. 14(1): 151.

Huotari T., Korpelainen H. (2012) *Complete chloroplast genome sequence of Elodea canadensis and comparative analyses with other monocot plastid genomes*. Gene 508(1): 96–105.

Kuang D.-Y., Wu H., Wang Y.-L., Gao L.-M., Zhang S.-Z., Lu L. (2011) *Complete chloroplast genome sequence of Magnolia kwangsiensis (Magnoliaceae): implication for DNA barcoding and population genetics*. Genome 54(8): 663–673.

Larson D.A., Walker J.F., Vargas O.M., Smith S.A. (2020) *A consensus phylogenomic approach highlights paleopolyploid and rapid radiation in the history of Ericales*. Am. J. Bot. 107(5): 773–789.

Miller M.A., Pfeiffer W., Schwartz T. (2010) *Creating the CIPRES Science Gateway for inference of large phylogenetic trees. 2010 Gateway Computing Environments Workshop (GCE)*. IEEE 2010: 1–8.

Nie X., Lv S., Zhang Y., Du X., Wang L., Biradar S.S., Tan X., Wan F., Weining S. (2012) *Complete chloroplast genome sequence of a major invasive species, crofton weed (Ageratina adenophora)*. PloSone 7(5): e36869.

Pauwels M., Vekemans X., Godé C., Frérot H., Castric V., Saumitou Laprade P. (2012) *Nuclear and chloroplast DNA phylogeography reveals vicariance among European populations of the model species for the study of metal tolerance, Arabidopsis halleri (Brassicaceae)*. New Phytologist 193(4): 916–928.

Perdereau A., Klaas M., Barth S., Hodkinson T.R. (2017) *Plastid genome sequencing reveals biogeographical structure and extensive population genetic variation in wild populations of Phalaris arundinacea L. in north western Europe*. Gcb Bioenergy 9(1): 46–56.

Posada D. (2008) *jModelTest: phylogenetic model averaging*. Mol. Biol. Evol. 25(7): 1253–1256.

Powell W., Morgante M., McDevitt R., Vendramin G., Rafalski J. (1995) *Polymorphic simple sequence repeat regions in chloroplast genomes: applications to the population genetics of pines*. Proc. Nat. Acad. Sci. 92(17): 7759–7763.

Provan J., Corbett G., Powell W., McNicol J. (1997) *Chloroplast DNA variability in wild and cultivated rice (Oryza spp.) revealed by polymorphic chloroplast simple sequence repeats*. Genome 40(1): 104–110.

Raubeson L.A., Peery R., Chumley T.W., Dziubek C., Fourcade H.M., Boore J.L., Jansen R.K. (2007) *Comparative chloroplast genomics: analyses including new sequences from the angiosperms Nuphar advena and Ranunculus macranthus*. BMC Gen. 8(1): 174.

Rose J.P., Kleist T.J., Löfstrand S.D., Drew B.T., Schönenberger J., Sytsma K.J. (2018) *Phylogeny, historical biogeography, and diversification of angiosperm order Ericales suggest ancient Neotropical and East Asian connections*. Mol. Phylogen. Evol. 122: 59–79.

Schönenberger J., Anderberg A.A., Sytsma K.J. (2005) *Molecular phylogenetics and patterns of floral evolution in the Ericales*. Int. J. Plant Sci. 166(2): 265–288.

Soltis P.S., Soltis D.E. (2021) *Plant genomes: markers of evolutionary history and drivers of evolutionary change*. Plants, People, Planet 3(1): 74–82.

Stamatakis A. (2014) *RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies*. Bioinformatics 30(9): 1312–1313.

Tong W., Kim T.-S., Park Y.-J. (2016) *Rice chloroplast genome variation architecture and phylogenetic dissection in diverse Oryza species assessed by whole-genome resequencing*. Rice 9(1): 57.

Yin K., Zhang Y., Li Y., Du F.K. (2018) *Different natural selection pressures on the atpF gene in evergreen sclerophyllous and deciduous oak species: evidence from comparative analysis of the complete chloroplast genome of Quercus aquifolioides with other oak species*. Int. J. Mol. Sci. 19(4): 1042.

Yuan C., Zhong W., Mou F., Gong Y., Pu D., Ji P., Huang H., Yang Z., Zhang C. (2017) *The complete chloroplast genome sequence and phylogenetic analysis of Chuanminshen (Chuanminshenviolaceum Sheh et Shan)*. Physiol. Mol. Biol. Plants 23(1): 35–41.

Zhang Y., Du L., Liu A., Chen J., Wu L., Hu W., Zhang W., Kim K., Lee S.-C., Yang T.-J. (2016) *The complete chloroplast genome sequences of five Epimedium species: lights into phylogenetic and taxonomic analyses*. Front. Plant Sci. 7: 306.

Zhao Y., Yin J., Guo H., Zhang Y., Xiao W., Sun C., Wu J., Qu X., Yu J., Wang X. (2015) *The complete chloroplast genome provides insight into the evolution and polymorphism of Panax ginseng*. Front. Plant Sci. 5: 696.