Int. J. Plant Sci. 181(8):000–000. 2020.

1058-5893/2020/18108-00XX$15.00 DOI: 10.1086/710083

THE CHLOROPLAST GENOME OF *SALVIA*: GENOMIC CHARACTERIZATION AND PHYLOGENETIC ANALYSIS

Fei Zhao,*,† Bryan T. Drew,‡ Ya-Ping Chen,* Guo-Xiong Hu,§ Bo Li,[1,]‖ and Chun-Lei Xiang[2,]*

*Key Laboratory for Plant Diversity and Biogeography of East Asia, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, Yunnan 650201, China; †University of Chinese Academy of Sciences, Beijing 100049, China; ‡Department of Biology, University of Nebraska–Kearney, Kearney, Nebraska 68849, USA; §College of Life Sciences, Guizhou University, Guiyang, Guizhou 550025, China; and ‖Research Centre of Ecological Sciences, College of Agronomy, Jiangxi Agricultural University, Nanchang 330045, China

Editor: Hervé Sauquet

Premise of research. Previous studies based on plastid fragments and/or nuclear ribosomal DNA have had limited success resolving relationships within the genus *Salvia*. This study evaluates the efficacy of complete plastome sequences for phylogenetic inference within *Salvia*, using the recently established *Salvia* subg. *Glutinaria* as a case study. We use these plastomes to identify hypervariable and simple sequence repeat (SSR) regions for future studies within *Salvia*.

Methodology. In order to produce a phylogenetic backbone for *Salvia*, we sequenced and assembled complete plastomes for six species of *Salvia*. These plastomes were combined with 11 plastomes (10 species) of *Salvia* from GenBank for analyses. This sampling represented seven of the 10 subgenera of *Salvia*. Genome features of these plastomes were analyzed, and hypervariable regions, SSRs, and longer repeats were identified. Phylogenetic relationships of 16 *Salvia* species were investigated using maximum likelihood and Bayesian methods based on four different data sets.

Pivotal results. All of the 17 *Salvia* plastomes displayed a typical quadripartite structure, and 114 different genes were identified in each accession. In addition, a total of 18 hypervariable regions and 626 SSRs were identified. The monophyly of *Salvia* and *Salvia* subg. *Glutinaria* was supported in our phylogenetic analyses.

Conclusions. Complete plastome sequences are promising for phylogenetic reconstruction of *Salvia* and likely other clades within Lamiaceae. In addition, we identified 18 hypervariable regions that should be useful as plastid phylogenetic markers for phylogenetic inferences within the genus and potentially as bar code markers for identifying different species of *Salvia*. The extended analysis of SSRs will be helpful for future population genetics studies and in elucidating the genetic diversity of *Salvia* and its relatives.

Keywords: Lamiaceae, Mentheae, Nepetoideae, Salviinae, *ycf1*.

Online enhancements: supplemental tables and figures.

Introduction

Salvia L. is the largest genus within Lamiaceae, with approximately 1000 species (Kriebel et al. 2019), and it includes well-known medicinal, ornamental, edible, and hallucinogenic plants (Will and Claßen-Bockhoff 2017). The large number of species, wide geographic breadth, and relatively high level of morphological variation within *Salvia* make this genus taxonomically challenging (González-Gallegos and Gama-Villanueva 2013; Will and Claßen-Bockhoff 2017; Kriebel et al. 2019; Drew et al. 2020; González-Gallegos et al. 2020). *Salvia* has a nearly worldwide distribution (Walker et al. 2004; Drew et al. 2017*a*; Will and Claßen-Bockhoff 2017; Hu et al. 2018; Kriebel et al. 2019) but has radiated extensively in three regions: Mesoamerica/South America, southwestern Asia and the Mediterranean region, and East Asia (Walker and Sytsma 2007). In East Asia, most species (ca. 83 out of ca. 100) are distributed and endemic to China (Li and Hedge 1994; Hu et al. 2014, 2017; Xiang et al. 2016; Ding et al. 2019). *Salvia* is one of the most medicinally important genera in China, and 45 species are used in traditional Chinese medicine (Peng and Xiang 2017). *Salvia* has traditionally been distinguished from other genera within Lamiaceae by a unique staminal morphology in which only two (as opposed to four in most of the related tribe Mentheae) functional stamens are expressed. Each of the two stamens has its thecae (anther sacs) separated via an elongated theca connective that is generally modified into the "staminal lever mechanism" (Claßen-Bockhoff et al. 2003).

[1] Author for correspondence; hanbolijx@163.com.
[2] Author for correspondence; xiangchunlei@mail.kib.ac.cn.

Manuscript received October 2019; revised manuscript received May 2020; electronically published September 2, 2020.

During the past two decades, substantial progress has been made toward clarifying phylogenetic relationships and generic boundaries within *Salvia* and related genera. Molecular phylogenetic studies showed that *Dorystaechas* Boiss. & Heldr. ex Benth., *Meriandra* Benth., *Perovskia* Kar., *Rosmarinus* L., and *Zhumeria* Rech. f. & Wendelbo are embedded within *Salvia* (Walker et al. 2004, 2015; Walker and Sytsma 2007; Drew and Sytsma 2011, 2012, 2013; Takano and Okada 2011; Jenks et al. 2013; Li et al. 2013; Will and Claßen-Bockhoff 2014, 2017; Drew et al. 2017*a*; Fragoso-Martínez et al. 2018; Kriebel et al. 2019). However, most of these five genera lack the elongated theca connective (and staminal lever mechanism) that has traditionally defined *Salvia*. As a result, there are two competing classifications of *Salvia* that have been proposed. The first is to synonymize the five embedded genera within *Salvia* (González-Gallegos 2015; Walker et al. 2015; Drew et al. 2017*a*, 2020; Hu et al. 2018, 2020; Kriebel et al. 2019), while the second proposal is to split *Salvia* into at least six genera (Will et al. 2015; Will and Claßen-Bockhoff 2017). On the basis of the reasons outlined by Drew et al. (2017*a*), a broadly defined *Salvia* is adopted here.

Investigations of the evolutionary history of *Salvia* and closely related genera using phylogenetic analysis have used a combination of 16 DNA markers (table 1; Walker et al. 2004, 2015; Taylor and Ayers 2006; Sudarmono and Okada 2007, 2008; Walker and Sytsma 2007; Jenks et al. 2011, 2013; Takano and Okada 2011; Drew and Sytsma 2012, 2013; Will and Claßen-Bockhoff 2014, 2017; Dizkirici et al. 2015; Will et al. 2015; Drew et al. 2017*a*, 2017*b*; Fragoso-Martínez et al. 2017, 2018; Hu et al. 2018). However, five (*trnT-trnL*, *trnV* intron, *trnS-trnG*, *atpB-rbcL*, and *rps16*) of these markers were used only once (Dizkirici et al. 2015; Walker et al. 2015) and were low in variability. For example, only one of 34 variable sites was parsimony informative within the *trnT-trnL* intergenic spacer (IGS) region (Dizkirici et al. 2015). In addition, although a backbone phylogeny was established on the basis of combined data sets using some of the aforementioned 16 DNA regions, resolution was poor for deep-level relationships (Walker and Sytsma 2007; Jenks et al. 2011, 2013; Takano and Okada 2011; Li et al. 2013; Will and Claßen-Bockhoff 2014, 2017; Walker et al. 2015; Fragoso-Martínez et al. 2018; Hu et al. 2018), and they were generally not useful in resolving relationships within smaller subclades. Development of more effective DNA markers is therefore necessary for resolving relationships within *Salvia*, and thus identification of hypervariable regions as candidates for phylogenetic study and identification of species of *Salvia* is needed.

Next-generation sequencing has provided a large amount of genome sequence data for green plants, and chloroplast genome sequencing has greatly contributed to elucidating phylogenetic and taxonomic problems as well as historical biogeographic inferences for flowering plants at different taxonomic levels (Jansen et al. 2007, 2011; Moore et al. 2007, 2010; Lin et al. 2010; Zhong et al. 2010; Xi et al. 2012; Barrett et al. 2014; Ma et al. 2014; Stull et al. 2015; Williams et al. 2016; Yu et al. 2017; Zhang et al. 2017*b*; Menezes et al. 2018; Xiang et al. 2020). A total of 51 complete chloroplast genomes from Lamiaceae have been reported. While *Salvia* is the largest genus in Lamiaceae, the complete chloroplast genome has been reported for only seven species (table S1; tables S1–S6 are available online; Qian et al. 2013; H. Chen et al. 2014; Ha et al. 2018; Y. P. Chen et al. 2019; Du et al. 2019; Liang et al. 2019). Consequently, little is known regarding structural variation within *Salvia* plastomes, and the phylogenetic utility of the complete plastome sequences within *Salvia* and tribe Mentheae remains unclear. Furthermore, the plastome can be a major resource for finding simple sequence repeats (SSRs), also known as microsatellites, which can be used

Table 1

DNA Markers Utilized in Previous Phylogenetic Studies of *Salvia*

DNA region	Reference(s)
rbcL	Walker et al. 2004; Sudarmono and Okada 2007, 2008; Takano and Okada 2011; Hu et al. 2018
trnL-trnF	Walker et al. 2004, 2015; Sudarmono and Okada 2007, 2008; Walker and Sytsma 2007; Jenks et al. 2011; Takano and Okada 2011; Drew and Sytsma 2012, 2013; Dizkirici et al. 2015; Will et al. 2015; Fragoso-Martínez et al. 2017, 2018; Hu et al. 2018
psbA-trnH	Walker and Sytsma 2007; Jenks et al. 2011, 2013; Walker et al. 2015; Fragoso-Martínez et al. 2017, 2018; Hu et al. 2018
ycf1	Drew and Sytsma 2012, 2013; Walker et al. 2015
ycf1-rps15	Drew and Sytsma 2012, 2013; Walker et al. 2015; Hu et al. 2018
rpl32-trnL	Drew and Sytsma 2012; Will and Claßen-Bockhoff 2014, 2017; Will et al. 2015
trnT-trnL	Dizkirici et al. 2015
trnV intron	Dizkirici et al. 2015
trnS-trnG	Walker et al. 2015
atpB-rbcL	Walker et al. 2015
rps16	Walker et al. 2015
trnK-matK	Walker et al. 2015
ITS	Taylor and Ayers 2006; Sudarmono and Okada 2007, 2008; Walker and Sytsma 2007; Jenks et al. 2011, 2013; Takano and Okada 2011; Drew and Sytsma 2012, 2013; Will and Claßen-Bockhoff 2014, 2017; Dizkirici et al. 2015; Walker et al. 2015; Will et al. 2015; Drew et al. 2017*b*; Fragoso-Martínez et al. 2017, 2018; Hu et al. 2018
ETS	Drew and Sytsma 2012; Will and Claßen-Bockhoff 2014, 2017; Walker et al. 2015; Will et al. 2015; Drew et al. 2017*b*; Hu et al. 2018
GBSSI	Drew and Sytsma 2012; Drew et al. 2017*a*
PPR-T3G09060	Drew and Sytsma 2012; Drew et al. 2017*a*

Note. ITS = internal transcribed spacer; ETS = external transcribed spacer.

as genetic markers for population genetics studies as they provide rich information for population genetics and evolutionary history (Powell et al. 1996). However, plastid SSRs have rarely been used within *Salvia* or Lamiaceae in general.

Although plastome phylogenies can be misleading because of issues associated with the maternal inheritance of the plastome, cpDNA sequences are nonetheless still important for phylogenetic studies of flowering plants, even in the face of hundreds of nuclear genes (e.g., Zeng et al. 2014; Couvreur et al. 2019; Zhang et al. 2019; reviewed by Jansen and Ruhlman 2012). In particular, cpDNA can be indispensable for detecting ancient and recent hybridization events, especially when used in concert with low-copy nuclear data. In this study, we analyzed plastomes from 17 accessions representing 16 species and 7 subgenera (out of 10) of *Salvia* for the following objectives: (1) present complete plastome sequences and compare plastid genomic structure and sequence variation within *Salvia*; (2) identify SSRs for potential use in future population genetics studies involving *Salvia*; (3) screen plastomes for hypervariable regions to use in phylogenetic analyses and species identification for future studies within *Salvia*; and (4) evaluate how effective whole plastomes are in phylogeny estimation within subclades of *Salvia*, using *Salvia* subg. *Glutinaria* as a case study.

Material and Methods

Taxon Sampling

In total, 23 species from subfamily Nepetoideae (Lamiaceae) were sampled. Ingroup taxa included 17 accessions representing 16 *Salvia* species: six species were newly sequenced, five species were reassembled on the basis of the whole-genome shotgun sequencing from the Sequence Read Archive (SRA) database, and others were downloaded from the National Center for Biotechnology Information database (table 2). This sampling represented seven of the 10 subgenera of *Salvia*. Our taxonomic sampling was mainly focused on the newly erected *Salvia* subg. *Glutinaria* (Hu et al. 2018) because we are interested in evaluating the utility of whole plastomes for future phylogenetic reconstruction of this subgenus. Published data on *S. japonica* (He et al. 2017) were not included for this analysis as we had doubts regarding its identification. Five species (*Dracocephalum palmatum* Stephan ex Willd., *Melissa officinalis* L., *Ocimum basilicum* L., *Origanum vulgare* L., and *Prunella vulgaris* L.) were selected as an outgroup for phylogenetic analyses (table 2). Voucher specimens for the newly sequenced taxa were deposited at the herbarium of the Kunming Institute of Botany (KUN), Chinese Academy of Sciences, and detailed information on the species sampled in the present study is provided in table 2.

DNA Extraction and Sequencing

Total genomic DNA was extracted from about 100 mg of fresh or silica gel–dried leaves using the modified CTAB method of Doyle and Doyle (1987), in which 4% CTAB was used with incorporation of 0.1% DL-dithiothreitol. The DNA concentration was measured using a NanoDrop 2000 spectrophotometer (Thermo Scientific, Carlsbad, CA) to ensure that the DNA concentration used for library construction was at least 30 ng/μL. Sample integrity and purification were detected by 1% agarose gel electrophoresis for 40 min at a voltage of 150 V.

The DNA samples (12 μg) were sheared into 300-bp fragments using a Covaris S2 instrument (Covaris). The fragmented DNA was combined with End-Repair Mix and incubated at 20°C for 30 min. The QIAquick PCR Purification Kit (Qiagen) was used to purify the end-repaired DNA, A-Tailing Mix was added, and it was incubated at 37°C for 30 min. The purified adenylate 3′-ends DNA was combined with adapter and ligation mix, and then the ligation reaction was incubated at 20°C for 15 min. Adapter-ligated DNA was selected by running a 2% agarose gel to recover the target fragments. Several rounds of PCR amplification with PCR primer cocktail and PCR master mix were performed to enrich the adapter-ligated DNA fragments. Last, the PCR products were selected by running another 2% agarose gel to recover the target fragments, and the gel was purified with the QIAquick Gel Extraction Kit (Qiagen). The final library was quantitated in two ways: (1) determining the average molecule length using the Agilent 2100 Bioanalyzer Instrument (Agilent DNA 1000 Reagents) and (2) quantifying the library with real-time quantitative PCR (TaqMan). The qualified libraries were first amplified within the flow cell on the cBot instrument for cluster generation (HiSeq 4000 PE Cluster Kit, Illumina). Then, the clustered flow cell was loaded onto the HiSeq 4000 sequencer for paired-end (PE) sequencing (HiSeq 4000 SBS Kit, Illumina) to generate PE 150-bp reads.

Genome Assembly and Annotation

Adapter sequences were trimmed, and low-quality reads were removed using fastq-mcf version 1.04.636 in the ea-utils package under the 64-bit Linux platform with default parameters (http://github.com/ExpressionAnalysis/ea-utils; Aronesty 2013). The quality of PE Illumina sequences was evaluated using the FastQC tool kit (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) with the parameter set as $Q \geq 25$ to acquire clean reads to ensure the high quality of the downstream analysis. After the raw data were filtered and low-quality reads were removed, high-quality PE reads were used for subsequent analyses. De novo assembling of the chloroplast genome was implemented in the GetOrganelle pipeline (http://github.com/Kinggerm/GetOrganelle; Jin et al. 2018). Contigs were connected with the help of Bandage version 0.8.1 (Wick et al. 2015) and were manually corrected when necessary. In order to validate the assembly error, we mapped the raw sequencing reads to the assembled plastid genome sequences using the Bowtie 2 (Langmead and Salzberg 2012) plug-in in Geneious version 11.0.3 (Kearse et al. 2012).

Chloroplast genome sequences were annotated using the online program Dual Organellar GenoMe Annotator with default parameters (Wyman et al. 2004). The putative starts, stops, and intron positions were manually adjusted according to comparisons with the plastome of *S. miltiorrhiza* in Geneious version 11.0.3 (Kearse et al. 2012). The tRNA boundaries were further verified using the online tRNAscan-SE service (Lowe and Chan 2016). The circular physical maps of all sequenced plastomes were drawn using the OrganellarGenomeDRAW tool (Lohse et al. 2013).

Simple and Long Repeat Sequence Analysis

The microsatellite search module MISA (Thiel et al. 2003), an SSR motif scanning tool written in Perl (http://pgrc.ipk-gatersleben.de/misa), was applied to identify SSRs in *Salvia* plastomes.

Table 2

Voucher Information and GenBank Accession Numbers for *Salvia* and Related Taxa Used in This Study

Taxon	Voucher, herbarium	Locality	GenBank no.	Reference
Ingroup:				
S. bulleyana Diels	na	na	MH603954	Liang et al. 2019
S. chanryoenica Nakai	s.n., KH	Mount Sobaek, South Korea	MH261357	Ha et al. 2018
S. digitaloides Diels	CL Xiang 1292, KUN	Lijiang, Yunnan, China	MN520016	This study
S. hispanica L.	na	na	MN520017	SRR6940040[a]
S. japonica Thunb.	Cultivated at Shaanxi Normal University	na	KY646143	He et al. 2017
S. meiliensis S.W. Su	GX Hu 0089, KUN	Yuexi, Anhui, China	MN520018	This study
S. miltiorrhiza Bunge	Cultivated at Institute of Medicinal Plant Development	Beijing, China	HF586694	Chen et al. 2014
S. nilotica Juss. ex Jacq.	GS Li 4276, IBSC	Kenya	MN520020	This study
S. officinalis L. 1	na	na	MN520021	SRR694004[a]
S. officinalis L. 2	na	na	MG772529	C. Schmiderer and J. Novak, unpublished data, 2018
S. petrophila G.X. Hu, E.D. Liu & Yan Liu	GX Hu 0292, KUN	Libo, Guizhou, China	MN520022	This study
S. przewalskii Maxim.	na	na	MH603953	Liang et al. 2019
S. rosmarinus Spenn.	Cultivated at Xi'an Botanical Garden	Xi'an, Shaanxi, China	KR232566	Chen and Hua 2019
S. sclarea L.	na	na	MN520023	SRR8534308[a]
S. merjamie Forssk.	GS Li 4279, IBSC	Kenya	MN520019	This study
S. splendens Sellow ex Wied-Neuw.	na	na	MN520024	SRR6382553[a]
S. yangii B.T. Drew	na	na	MN520025	SRR6940082[a]
S. yunnanensis C.H. Wright	GX Hu QT001, KUN	Kunming, Yunnan, China	MN520026	This study
Outgroup:				
Dracocephalum palmatum Stephan ex Willd.	na	na	KU958581	M. R. Kabilov et al., unpublished data, 2016
Melissa officinalis L.	na	na		SRR6940050[a]
Ocimum basilicum L.	C. Lee 007	na	KY623639	Rabah et al. 2017
Origanum vulgare L. ssp. *vulgare*	University of Veterinary Medicine, Vienna (accession SR1985)	Bad Deutsch-Altenburg, Austria	JX880022	Lukas and Novak 2013
Prunella vulgaris L.	University of Macau (accession HYW170001)	Zhejiang, China	MG589640	Han and Zheng 2018

Note. na = not available.
[a] Sequence Read Archive.

Thresholds for a minimum number of repeat units were 10 for mononucleotides; five for dinucleotides; four for trinucleotides; and three for tetranucleotides, pentanucleotides, and/or hexanucleotides. Four types of long repeat sequences in *Salvia* plastomes, direct (forward), inverted (palindromic), complement, and reverse repeats, were determined using the online REPuter program (Kurtz et al. 2001) according to the following criteria: cutoff ≥30 bp and 90% sequence identities (Hamming distance of 3).

Chloroplast Genome Comparison and Sequence Divergence Analysis

Published complete plastome sequences of *Salvia* provide an opportunity to compare the sequence variation within the genus. The software mVISTA (Frazer et al. 2004) was used to evaluate variability of the complete plastome sequences among the *Salvia* species, default parameters were used to align the plastomes under LAGAN mode, and *S. przewalskii* (MH603953) was used as the reference. The Mauve version 2.3.1 (Darling et al. 2004) plug-in in Geneious version 11.0.3 (Kearse et al. 2012) was used with the default parameters to identify locally collinear blocks among the 17 *Salvia* plastomes.

All the sequences of *Salvia* were aligned using MAFFT version 7.271 (Katoh and Standley 2013) and were manually adjusted in Geneious version 11.0.3 (Kearse et al. 2012). The nucleotide diversity (π) of the chloroplast genome was estimated in a sliding-window analysis using DnaSP version 6.0 (Rozas et al. 2017). The step size was set to 200 bp, with a 600-bp window length.

Codon Usage Analysis

The relative synonymous codon usage (RSCU) is the ratio between the frequency of use and the expected frequency of a particular codon and is a simple measure of nonuniform usage of synonymous codons in a coding sequence. Here, the

codon usage was determined for all protein-coding genes. The software DAMBE (Xia and Xie 2001) was used to calculate the value of RSCU, and the histogram was drawn using the R package ggplot2 (Wickham 2009).

Phylogenetic Analysis

In this study, four data sets were generated for phylogenetic inference. The first data set consisted of 22 complete plastid genome sequences (CPGs) with one inverted repeat (IR) region excluded, the second data set contained 80 protein-coding genes and named coding regions (CRs), the third data set included the IGSs and introns (noncoding region [NCR]), and the fourth data set included only the 18 hypervariable regions. Because our goal with the fourth data set was to assess the utility and effectiveness of the hypervariable regions for phylogenetic study within *Salvia*, we used only *M. officinalis* as an outgroup (on the basis of Drew and Sytsma 2012). Before the establishment of the combined matrices, each protein-coding gene, IGS, and intron was realigned using MAFFT version 7.271 (Katoh and Standley 2013) with the L-INS-i algorithm.

Maximum likelihood (ML) analyses were conducted using RAxML version 8.1.11 (Stamatakis 2014) as implemented on the Cyberinfrastructure for Phylogenetic Research (CIPRES) Science Gateway (http://www.phylo.org/; Miller et al. 2010), with 1000 bootstrap iterations (−#|−N) and other parameters using the default settings. Bayesian inference (BI) analyses were carried out with MrBayes version 3.2.3 (Ronquist et al. 2012) as implemented on CIPRES using the default settings. Markov chain Monte Carlo analysis was executed for 20 million generations with four chains (one cold and three heated), each starting with a random tree and sampled every one-thousandth generation. Convergence of runs was accepted when the average standard deviation of split frequencies dropped below 0.01. Tracer version 1.6.0 (Rambaut et al. 2014) was used to check that the effective sample size values were ≥200. The first 25% of the trees were discarded as burn-in, and the remaining trees were used to construct majority-rule consensus trees.

Phylogenetic trees were edited using TreeGraph 2 (Stöver and Müller 2010), and we defined branches with posterior probabilities (PP) < 0.90 and bootstrap values (BS) < 70% as weakly supported, with PP = 0.90 and BS = 70%–80% as moderately supported, and with PP ≥ 0.95 and BS ≥ 80% as strongly supported (Chen et al. 2019).

Results

Genome Assembly and Chloroplast Genome Characterization of Salvia

For the six newly sequenced species, Illumina PE sequencing generated 52,378,453 (*S. merjamie* Forssk.) to 68,112,046 (*S. meiliensis* S. W. Su) clean reads, with mean coverage from 837 (×) in *S. nilotica* to 1970 (×) in *S. meiliensis*. For the five species reassembled on the basis of the whole-genome shotgun sequencing from the SRA database, the PE reads ranged from 16,847,780 (*S. sclarea* L.) to 128,316,200 (*S. splendens* Sellow ex J.A. Schultes), with mean coverage from 362 (×) in *S. sclarea* to 1944 (×) in *S. splendens*. Among the 17 *Salvia* plastomes, genome size ranged from 150,604 bp in *S. splendens* to 152,462 bp in *S. rosmarinus* (table 3).

All of the 17 *Salvia* plastomes displayed a quadripartite structure that consisted of a pair of IR regions (25,283–25,623 bp) separated by the large single-copy (LSC; 82,181–83,400 bp) and small single-copy (SSC; 17,464–17,967 bp) regions (table 3); thus, only the chloroplast genome maps of *S. merjamie* and *S.*

Table 3

Accession Numbers and Features of the 17 *Salvia* Plastomes in the Present Study

Taxon	Accession no.	Clean reads	Reads used in assembly	Mean coverage of base (×)	Complete		LSC		SSC		IR	
					Length (bp)	GC (%)	Length (bp)	GC (%)	Length (bp)	GC (%)	Length (bp)	GC (%)
S. bulleyana	MH603954	na	na	na	151,547	38.00	82,853	36.10	17,592	31.90	25,551	43.10
S. chanryoenica	MH261357	na	na	na	151,689	38.00	82,903	36.10	17,630	32.00	25,578	43.10
S. hispanica	MN520017	33,007,821	28,430,914	1127	150,980	38.00	82,279	36.20	17,535	31.60	25,583	43.10
S. miltiorrhiza	HF586694	na	na	na	151,332	38.00	82,689	36.20	17,556	32.00	25,539	43.10
S. officinalis 1	MN520021	25,242,685	18,665,433	1075	151,135	38.00	82,463	36.20	17,500	31.90	25,586	43.10
S. officinalis 2	MG772529	na	na	na	151,089	38.00	82,407	36.20	17,500	31.90	25,591	43.10
S. przewalskii	MH603953	na	na	na	151,319	38.00	82,732	36.10	17,605	31.90	25,491	43.10
S. rosmarinus	KR232566	na	na	na	152,462	38.00	83,355	36.20	17,967	31.90	25,571	43.10
S. sclarea	MN520023	16,847,780	12,173,289	362	151,268	37.90	82,535	36.10	17,595	31.70	25,569	43.10
S. splendens	MN520024	128,316,200	17,751,167	1944	150,604	38.00	82,181	36.20	17,857	31.90	25,283	43.20
S. yangii	MN520025	31,776,790	8,725,805	854	151,466	38.10	82,694	36.20	17,566	31.90	25,603	43.10
S. yunnanensis	MN520026	67,605,182	23,963,346	1564	151,413	38.00	82,656	36.10	17,577	32.00	25,590	43.10
S. digitaloides	MN520016	67,784,948	26,595,742	853	152,159	38.00	83,400	36.20	17,579	31.90	25,590	43.10
S. petrophila	MN520022	67,976,942	27,112,112	932	151,688	38.00	82,883	36.10	17,641	31.80	25,582	43.10
S. nilotica	MN520020	56,097,063	26,694,277	837	151,450	38.00	82,740	36.10	17,464	31.90	25,623	43.10
S. meiliensis	MN520018	68,112,046	25,577,470	1970	151,614	38.00	82,854	36.10	17,580	32.00	25,590	43.10
S. merjamie	MN520019	52,378,453	25,467,727	1226	151,286	37.90	82,566	36.10	17,586	31.70	25,567	43.10

Note. na = not available; GC = guanine-cytosine; LSC = large single copy; SSC = small single copy; IR = inverted repeat.

digitaloides are presented as representatives of *Salvia* (fig. 1), and others are provided in the supplemental material (fig. S1; figs. S1–S5 are available online).

The guanine-cytosine (GC) content was evenly distributed, and the average GC content was 38.0% (table 3). In general, the GC content in the IR regions (43.1%–43.2%) was higher than in the LSC (36.1%–36.2%) and SSC (31.6%–32.0%) regions. Accession numbers of all newly sequenced plastomes are listed in table 3.

Seventeen genes were duplicated in the IR region, with six protein-coding genes, seven tRNAs, and four rRNAs (table 4). When duplicated genes in the IR regions were counted only once, each plastome harbored 114 different genes, including 80

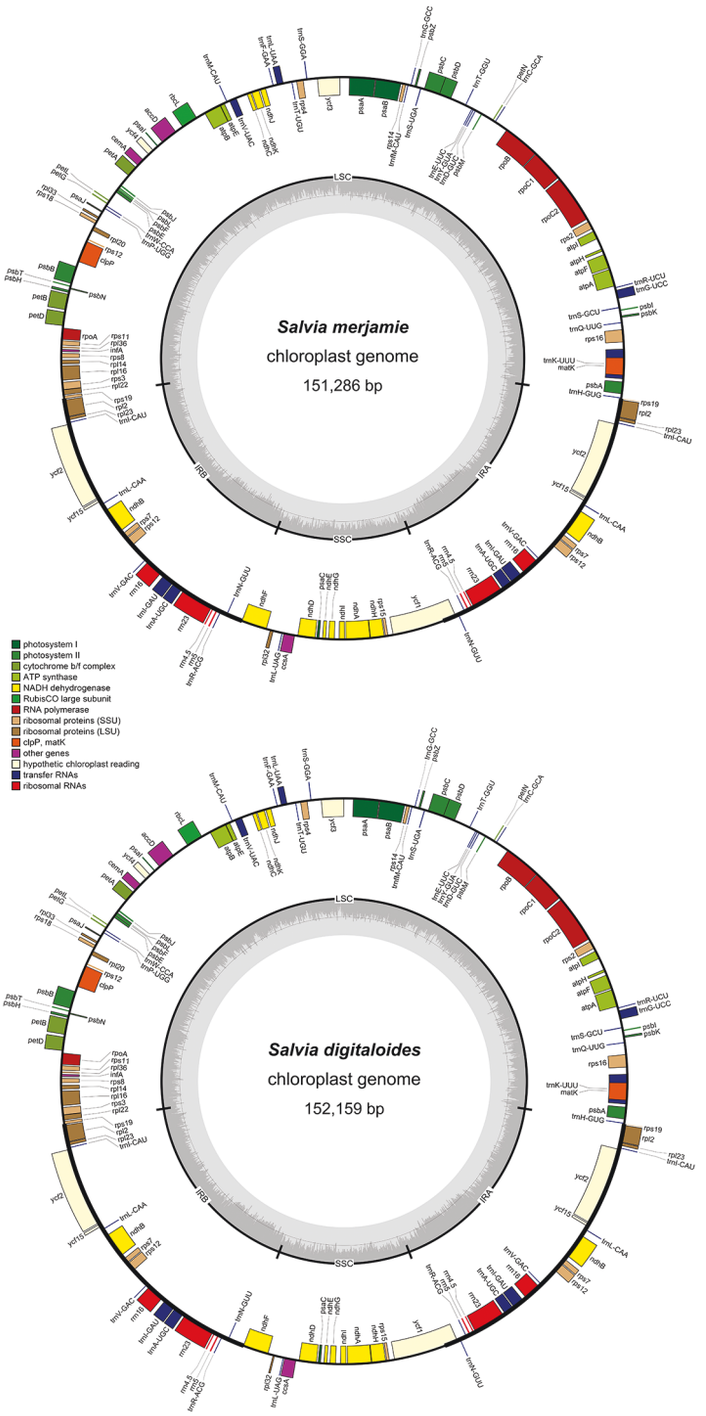

Fig. 1 Gene map of chloroplast genomes of *Salvia merjamie* and *S. digitaloides*. Genes inside and outside the circle are transcribed in the clockwise and counterclockwise directions, respectively. Genes belonging to different functional categories are color-coded.

Table 4

Gene Contents and Functions of *Salvia*

Category of genes, group of genes	Name(s) of genes
Self-replication:	
Ribosomal RNA genes	*rrn16*,[c] *rrn23*,[c] *rrn4.5*,[c] *rrn5*[c]
Transfer RNA genes, 30 tRNA genes	6 contain an intron; 7 are duplicated in the IR region; *trnA-UGC*,[a,c] *trnfM-CAU*, *trnI-GAU*,[a,c], *trnM-CAU*, *trnR-ACG*,[c] *trnS-UGA*, *trnC-GCA*, *trnG-GCC*,[a] *trnK-UUU*,[a] *trnN-GUU*,[c] *trnW-CCA*, *trnT-GGU*, *trnD-GUC*, *trnG-UCC*, *trnL-CAA*,[c] *trnY-GUA*, *trnR-UCU*, *trnT-UGU*, *trnE-UUC*, *trnH-GUG*, *trnL-UAA*,[a] *trnP-UG*, *trnS-GCU*, *trnV-GAC*,[c] *trnF-GAA*, *trnI-CAU*,[c] *trnL-UAG*, *trnQ-UUG*, *trnS-GGA*, *trnV-UAC*[a]
Small subunit of ribosome	*rps2*, *rps3*, *rps4*, *rps7*,[c] *rps8*, *rps11*, *rps12*, *rps14*, *rps15*, *rps16*,[a] *rps18*, *rps19*
Large subunit of ribosome	*rpl2*,[a,c] *rpl14*, *rpl16*,[a] *rpl20*, *rpl22*, *rpl23*,[c] *rpl32*, *rpl33*, *rpl36*
RNA polymerase subunits	*rpoA*, *rpoB*, *rpoC1*,[a] *rpoC2*
Photosynthesis:	
Subunits of NADH dehydrogenase	*ndhA*,[a] *ndhB*,[a,c] *ndhC*, *ndhD*, *ndhE*, *ndhF*, *ndhG*, *ndhH*, *ndhI*, *ndhJ*, *ndhK*
Photosystem I	*psaA*, *psaB*, *psaC*, *psaI*, *psaJ*, *ycf3*[b]
Photosystem II	*psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbJ*, *psbK*, *psbL*, *psbM*, *psbN*
Cytochrome b/f complex	*petA*, *petB*,[a] *petD*,[a] *petG*, *petL*, *pet*
ATP synthase	*atpA*, *atpB*, *atpE*, *atpF*,[a] *atpH*, *atpI*
Large chain of Rubisco	*rbcL*
Other genes:	
Translation initiation factor	*infA*
Maturase	*matK*
Protease	*clpP*[b]
Envelope membrane protein	*cemA*
Subunit of acetyl-CoA-carboxylase	*accD*
Cytochrome c biogenesis protein	*ccsA*
Component of TIC complex	*ycf1*
Genes of unknown functions	*ycf2*, *ycf4*, *ycf15*[c]

Note. IR = inverted repeat.

[a] Gene with a single intron.

[b] Gene with two introns.

[c] Gene with duplicated introns.

protein-coding genes, 30 tRNAs, and 4 rRNAs (tables 4, 5), that were arranged in the same order. Nine protein-coding genes and seven tRNA genes contained one intron, and three genes (*clpP*, *rps12*, and *ycf3*) contained two introns (table 4). Within those sampled sequences in the study, protein-coding regions accounted for 51.11%–53.45% of the whole genome, while tRNA and rRNA regions accounted for 1.83%–1.86% and 5.94%–6.01%, respectively (table 5). The remaining regions were noncoding sequences, including IGSs, introns, and pseudogenes.

SSR Polymorphisms and Long Repeat Sequence Analysis

In total, 626 SSRs were detected among the 17 *Salvia* plastomes (fig. 2; table S2), with length variation from 10 to 188 bp. Comparative analysis shows that all six types of SSRs (mononucleotide, dinucleotide, trinucleotide, tetranucleotide, pentanucleotide, and hexanucleotide repeats) were detected, while the number of each repeat unit was different. Among these SSRs, the majority consisted of mononucleotide, dinucleotide, and tetranucleotide repeats, which were found 419, 81, and 118 times, respectively. In contrast, the tri-, penta-, and hexanucleotides showed a lower frequency, appearing only three, two, and three times, respectively. In the mononucleotide repeat unit, the proportions of mononucleotide repeat unit A/T and C/G repeats were 98.57% and 1.43%, respectively. In the dinucleotide repeat unit, 66.7% of the dinucleotide repeat sequences consisted of AT/AT repeats, and the remainders were AC/GT and AG/CT repeats. The trinucleotide repeat unit (AAT/ATT) was detected in *S. yunnanensis*, *S. sclarea*, and *S. hispanica*. The tetranucleotide repeat unit included seven different sequence types, and the sequence repeat units AAAT/ATTT and ACAG/CTGT were shared by all 17 plastomes. The other five tetranucleotide repeat types were uncommon and were observed in only a few species. The pentanucleotide repeat sequence (AATCT/AGATT) was found only in *S. przewalskii*. The hexanucleotide repeat unit contained three different types (AAAATC/ATTTTG, AATTAT/AATTAT, and AAGTCT/ACTTAG), and these were detected only once in *S. meiliensis*, *S. yangii*, and *S. miltiorrhiza*, respectively (table 6). Within the 17 *Salvia* plastomes, SSR loci were mainly located in the LSC region (83.39%; fig. 2), followed by the SSC region (13.42%) and the IR regions (3.19%).

A total of 715 long repeat sequences, including forward, reverse, palindromic, and complementary repeats, were detected in the 17 *Salvia* plastomes (fig. 3; table S3). The number of repeats ranged from 30 (*S. merjamie* and *S. sclarea*) to 49 (*S. hispanica*, *S. miltiorrhiza*, *S. petrophila*, *S. rosmarinus*, and *S. splendens*), and the size ranged from 30 to 216 bp. Most repeats were shorter than 100 bp, while four forward repeats longer than 100 bp were found in *S. rosmarinus*. The forward repeat, which accounted for 48.95% of the total repeats, was the most frequent type, followed by palindromic repeats (48.11%), reverse repeats (1.96%), and complementary repeats (0.98%). Reverse repeats

Table 5

Gene Numbers and Length Percentages of Protein Coding, rRNA, and tRNA in *Salvia* Plastomes

Taxon	Total genes	Protein-coding genes	tRNA genes	rRNA genes	Protein coding		rRNA sequence		tRNA sequence	
					Length (bp)	Percentage (%)	Length (bp)	Percentage (%)	Length (bp)	Percentage (%)
S. bulleyana	114	80	30	4	79,335	52.35	9064	5.98	2791	1.84
S. chanryoenica	114	80	30	4	80,076	52.78	9052	5.97	2797	1.84
S. digitaloides	114	80	30	4	80,715	53.05	9048	5.95	2789	1.83
S. hispanica	114	80	30	4	79,419	52.60	9052	6.00	2787	1.85
S. meiliensis	114	80	30	4	79,186	52.23	9048	5.97	2794	1.84
S. merjamie	114	80	30	4	80,856	53.45	9052	5.98	2789	1.84
S. miltiorrhiza	114	80	30	4	80,722	53.34	9084	6.00	2818	1.86
S. nilotica	114	80	30	4	80,802	53.35	9048	5.97	2789	1.84
S. officinalis 1	114	80	30	4	79,053	52.31	9048	5.99	2789	1.85
S. officinalis 2	114	80	30	4	79,389	52.54	9052	5.99	2792	1.85
S. petrophila	114	80	30	4	79,401	52.34	9048	5.96	2789	1.84
S. przewalskii	114	80	30	4	79,449	52.50	9052	5.98	2791	1.84
S. rosmarinus	114	80	30	4	77,925	51.11	9056	5.94	2794	1.83
S. sclarea	114	80	30	4	79,494	52.55	9052	5.98	2789	1.84
S. splendens	114	80	30	4	79,011	52.46	9048	6.01	2789	1.85
S. yangii	114	80	30	4	80,609	53.22	9052	5.98	2790	1.84
S. yunnanensis	114	80	30	4	79,458	52.48	9046	5.97	2789	1.84

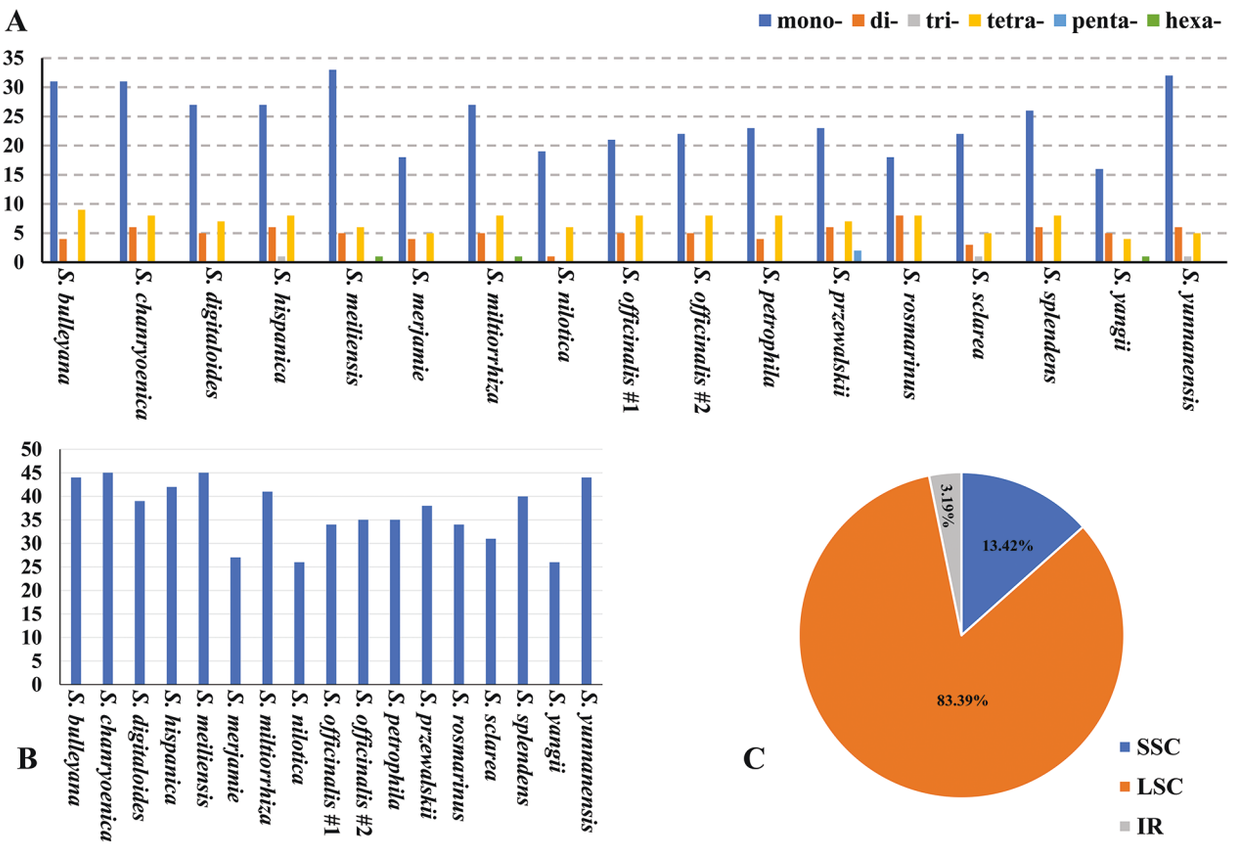

Fig. 2 Numbers, types, and distribution of simple sequence repeats (SSRs) in the 17 *Salvia* plastomes. *A*, Numbers and types of SSRs. *B*, Total number of SSRs detected in each species. C, Presence of SSRs in the large single-copy (LSC), small single-copy (SSC), and inverted repeat (IR) regions.

Table 6

Numbers and Types of the Simple Sequence Repeats of the 17 *Salvia* Plastomes

Type, repeat sequence	*S. bull*	*S. chan*	*S. digi*	*S. hisp*	*S. meil*	*S. merj*	*S. milt*	*S. nilo*	*S. offi* 1	*S. offi* 2	*S. petr*	*S. prze*	*S. rosm*	*S. scla*	*S. sple*	*S. yang*	*S. yunn*	Percentage (%)
Mono:																		
A/T	31	31	27	26	32	18	27	19	21	22	23	23	17	22	26	16	32	66.93
C/G	1	1	1	1	1								1					
Di:																		
AC/GT	3	5	4										1					
AG/CT				1	1	1	1		1	1	1	1	2	1	1	1	1	12.94
AT/AT				5	4	3	4	1	4	4	3	5	5	2	5	4	5	
Tri:																		
AAT/ATT				1										1			1	.48
Tetra:																		
AAAC/GTTT	1	1	1	1	1		1	1	2	2	2	1			1	1	1	
AAAG/CTTT	1	1	1	1	1	1	1	2	2	2	1	1	2	1	1			
AAAT/ATTT	4	4	4	4	3	2	4	1	2	2	3	3	3	2	4	2	3	
AAAT/ATTT													1					18.85
AATT/AATT	1						1											
ACAG/CTGT	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
AGAT/ATCT	1	1		1		1		1	1	1	1	1	1	1	1			
Penta:																		
AATCT/AGATT												2						.32
Hexa:																		
AAAATC/ATTTTG					1													
AATTAT/AATTAT																1		.48
AAGTCT/ACTTAG							1											
Total	44	45	39	42	45	27	41	26	34	35	35	38	34	31	40	26	44	100

Note. *S. bull* = *S. bulleyana*; *S. chan* = *S. chanryoenica*; *S. digi* = *S. digitaloides*; *S. hisp* = *S. hispanica*; *S. meil* = *S. meiliensis*; *S. merj* = *S. merjamie*; *S. milt* = *S. miltiorrhiza*; *S. nilo* = *S. nilotica*; *S. offi* 1 = *S. officinalis* 1; *S. offi* 2 = *S. officinalis* 2; *S. petr* = *S. petrophila*; *S. prze* = *S. przewalskii*; *S. rosm* = *S. rosmarinus*; *S. scla* = *S. sclarea*; *S. sple* = *S. splendens*; *S. yang* = *S. yangii*; *S. yunn* = *S. yunnanensis*.

were detected only in *S. merjamie* (one), *S. rosmarinus* (12), and *S. sclarea* (one). The complementary repeats were found in *S. meiliensis* and *S. rosmarinus*, with frequencies of one and six, respectively. These repeats were scattered around the complete plastomes; 474 (66.3%) were located in the IR regions, and 241 (33.7%) were located in the single-copy region. The majority (56.6%) of those repeats were located in CRs (i.e., *accD*, *psaB*, *ycf1*, and *ycf2*), and the others were found in intronic regions (17.3%) and intergenic regions (26.1%).

Comparative Analysis of the Chloroplast Genome of Salvia

All the genes within the *Salvia* plastomes were found in the same order, and the IRa and IRb regions were more conserved than the LSC and SSC regions (fig. 4). In addition, the CRs presented higher sequence identity than the NCRs, and the most divergent regions among the chloroplast genomes occurred within IGSs (i.e., *rps16-trnQ* and *rpl32-trnL*). The variation of the IR regions among all *Salvia* species included for analyses was small and mostly conserved. Except for *S. hispanica*, the LSC/IRa junction of the other *Salvia* taxa was located within the *rps19* gene, ranging from 24 to 45 bp away from the LSC/IRa junction, which caused a pseudogene fragment (*ψrps19*) at the IRb/LSC border. For most species, the IRa region expanded into *ndhF*, with the length ranging from 17 to 36 bp (fig. 5). However, the IRa/SSC junction in *S. rosmarinus* was distinctly different compared with those of the other species in this study. In *S. rosmarinus*, the *ndhF* gene was completely located within the SSC region at a distance of 213 bp from the IRa/SSC border. In all species, the *ycf1* gene crosses the SSC/IRb region, with a length variation from 864 bp to 1171 bp and an equal length of pseudogene *ycf1* (*ψycf1*) detected in the IRa/SSC boundary. In addition, the *trnH* gene is separated from the IRa/LSC border by a spacer varying from 3 to 65 bp.

The results of the sequence divergence analyses of the 17 *Salvia* plastomes showed that the π values ranged from 0 (*rrn16* gene) to 0.066 (*ycf1* gene), with an average of 0.0136 (fig. 6; table S4). In total, 18 regions with high divergence values ($\pi > 0.03$) were detected in the 17 plastomes (fig. 6); these are termed hypervariable regions here. Among these 18 regions, 11 of those loci were located in the LSC region, seven in the SSC region, and none in

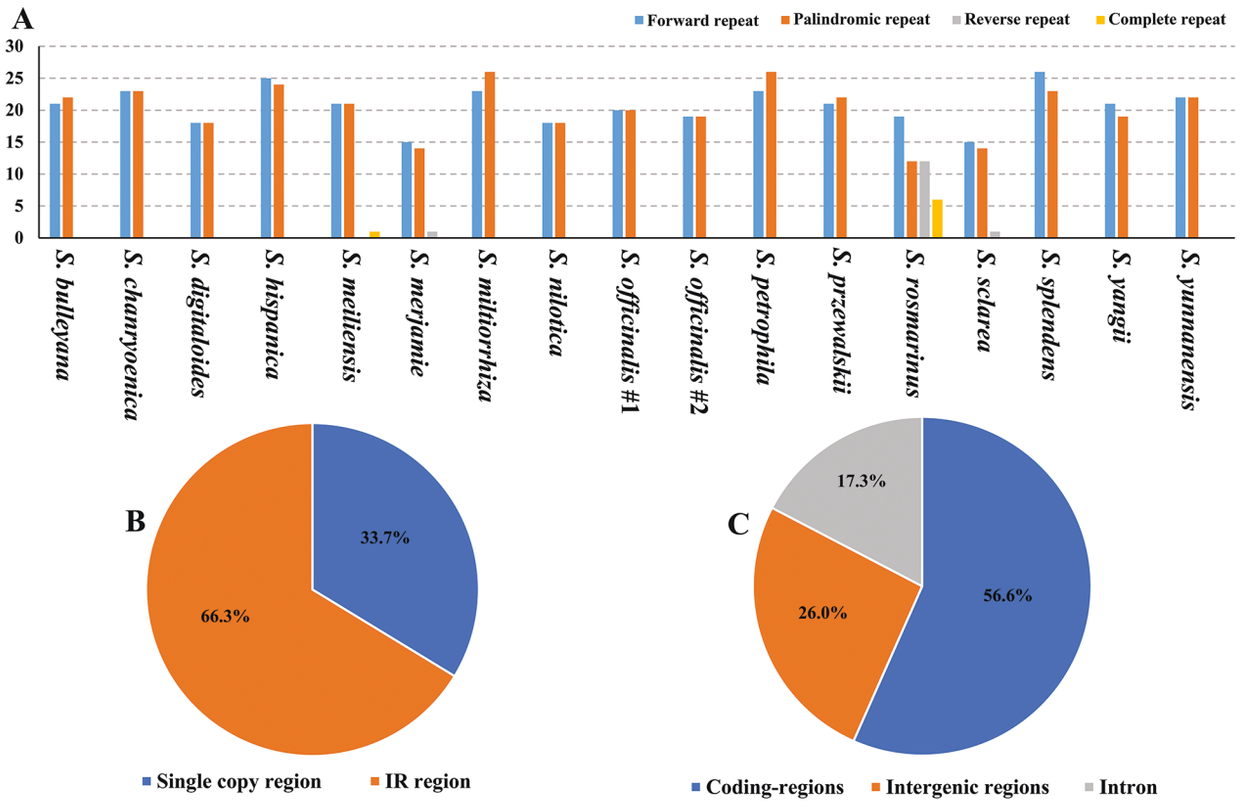

Fig. 3 Numbers, types, and distribution of long repeat sequences in the 17 *Salvia* plastomes. *A*, Numbers and types of longer repeats. *B*, Presence of longer repeats in the single-copy region and the inverted repeat (IR) region. *C*, Presence of longer repeats in protein-coding regions, intergenic regions, and intron regions.

the IR region. Fifteen of the hypervariable regions were intergenic regions or introns (*trnK-rps16*, *rps16* intron-*trnQ*, *psbK-psbC*, *atpH-atpI*, *rpoB-trnC-petN*, *trnE-trnT*, *rbcL-accD*, *petA-pabJ*, *rpl16* intron, *rps3-rpl22*, *rpl32-trnL-ccsA*, *ccsA-ndhD*, *ndhG-ndhI*, *ndhA* intron, and *rps15-ycf1*), while three were protein-coding genes (*matK*, *ndhF*, and *ycf1*).

Codon Usage

The length of the concatenated data set of 80 protein-coding genes from the 17 *Salvia* plastomes ranged from 68,763 to 69,009 bp and was composed of between 22,921 and 23,003 codons. The codon content for the 20 amino acids and the stop codon for the 80 protein-coding genes of the *Salvia* plastomes are summarized in figure 7 and table S5. Among these codons, the amino acid in the highest proportion was leucine (10.61%), followed by isoleucine (8.45%) and serine (7.59%), while cysteine was the least frequent amino acid, with a percentage of 1.94%. The standard initiator codon AUG was employed by most protein-coding genes; however, other start codons were identified in the *rps19* gene (GTG) and the *ndhD* gene (TTG, CTG). The RSCU value analysis showed that, except for methionine and tryptophan, other amino acids have more than one synonymous codon. In *Salvia* plastomes, the codons of chloroplast genes with A/U at the third-position nucleotide were more frequent than those ending with G/C, according to RSCU values (with a threshold of RSCU > 1).

Phylogenetic Analysis

Properties and statistical characters of the four data sets used in this study are summarized in table 7. The aligned CPG data set had a length of 134,646 bp, of which 117,632 bp (87.36%) were constant and 6597 bp (4.90%) were parsimony informative. The aligned data sets of the combined CRs and combined NCRs were 68,987 bp (CR data set) and 61,615 bp (NCR data set) and consisted of 2681 bp (3.89%) and 3868 bp (6.28%) of parsimony-informative sites, respectively. The combined 18 hypervariable regions had an aligned length of 26,028 bp, with 22,515 constant sites (86.50%) and 1774 parsimony-informative sites (6.82%). As expected, the combined 18 hypervariable regions data set showed the highest percentage of parsimony-informative sites (6.82%) among those data sets, while the CR data set (3.89%) showed the lowest.

The backbones of the trees constructed from ML and BI analyses of each of the four combined data sets were identical (figs. S2–S4). The tree from the ML analysis of the CR data set is the representative tree illustrated for discussion of phylogenetic relationships (fig. 8).

In all analyses, the monophyly of *Salvia* was recovered with strong support (fig. 8; ML BS: 100%; BI PP: 1.00; all values reported in this order below). Representatives of subgenera *Perovskia*, *Rosmarinus*, *Salvia*, and *Sclarea* and the "*Heterosphace*" clade formed a clade sister to subgenera *Calosphace* and *Glutinaria*. In this clade, the first and second branching members

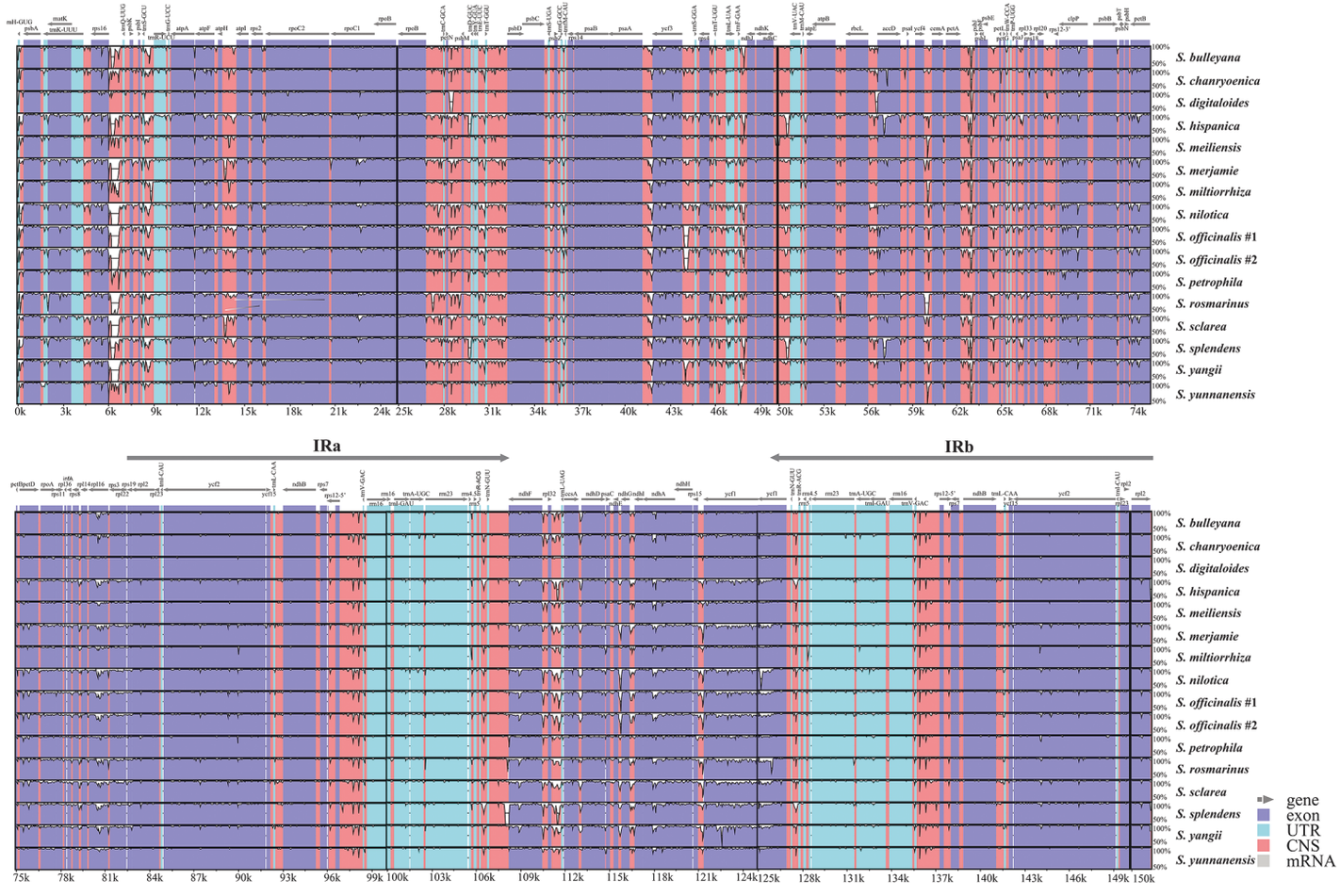

Fig. 4 Comparisons of percentage identity of chloroplast genomes of *Salvia*. *Salvia przewalskii* was used as a reference. Genome regions are color-coded as exons, untranslated regions (UTRs), conserved noncoding sequences (CNSs), and mRNA. The vertical scale indicates the percent identity, ranging from 50% to 100%. IR = inverted repeat.

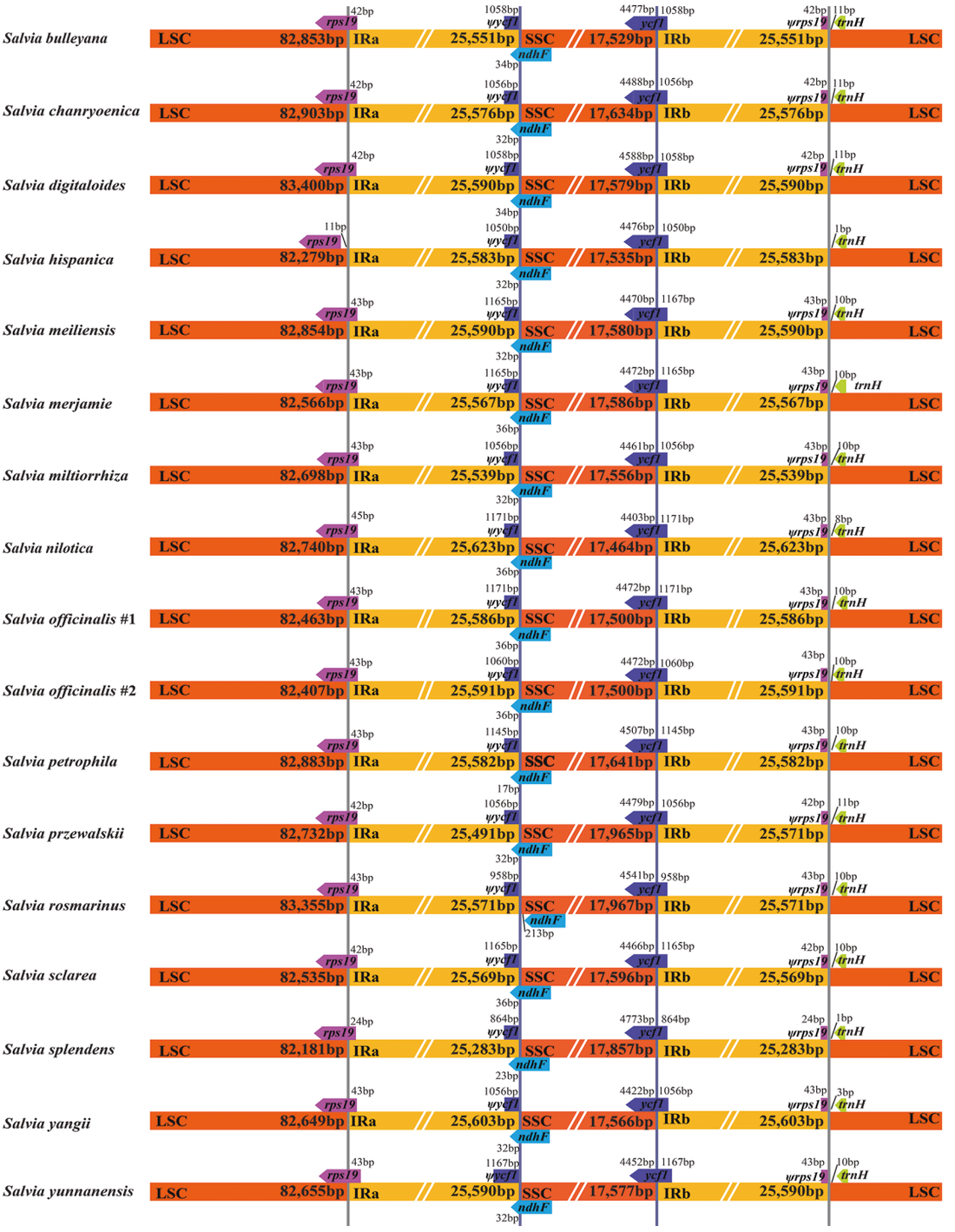

Fig. 5 Comparison of junctions between the inverted repeat (IR), large single-copy (LSC), and small single-copy (SSC) regions (orange, yellow, and green areas, respectively) of the 17 *Salvia* genomes. The distance between the gene and the boundary is not proportionate. The ψ notation indicates a pseudogene.

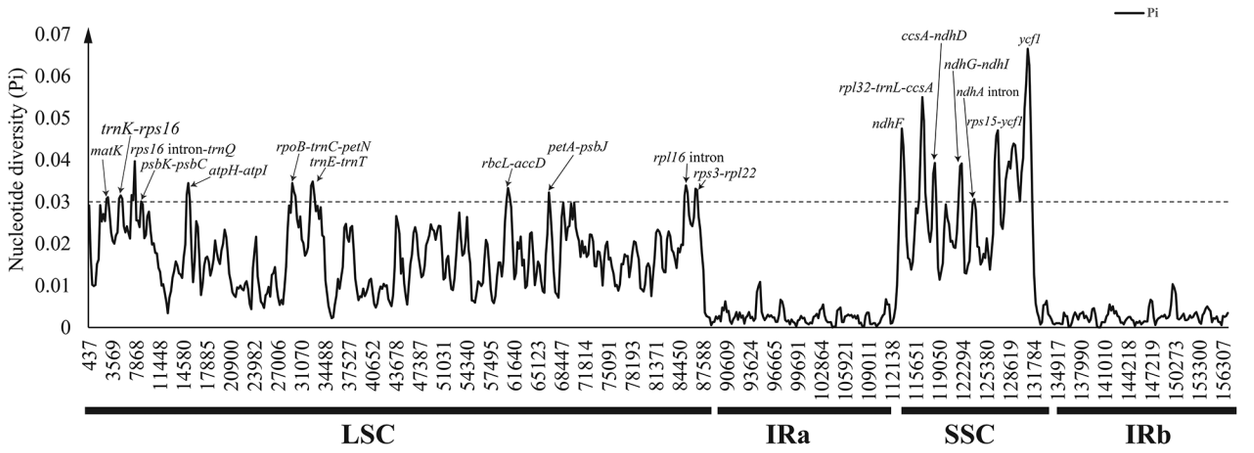

Fig. 6 Sliding-window analysis of the 17 *Salvia* plastomes. The *X*-axis indicates the position of the midpoint of a window; the *Y*-axis indicates the nucleotide diversity of each window. The 18 hypervariable regions are marked by the arrows. LSC = large single copy; IR = inverted repeat; SSC = small single copy.

were *S. rosmarinus* (subg. *Rosmarinus*) and *S. yangii* (subg. *Perovskia*), which were sister to *S. nilotica* ("*Heterosphace*" clade). The first was sister to a clade of two individuals of *S. officinalis*, *S. sclarea*, and *S. merjamie* (100%; 1.00). These four species formed part of a large clade in previous studies (e.g., clade I in Will and Claßen-Bockhoff 2014, 2017).

Two species (*S. splendens* and *S. hispanica*) of *Salvia* subg. *Calosphace* were recovered as a clade sister to *Salvia* subg. *Glutinaria* (100%; 1.00), consisting of eight species. Within *Salvia* subg. *Glutinaria*, as reported by Hu et al. (2018), *S. petrophila* of sect. *Sonchifoliae* was sister to all remaining taxa. *Salvia chanryoenica* (sect. *Glutinaria*) diverged next, followed by sect. *Eurysphace* (*S. digitaloides*, *S. przewalskii*, and *S. bulleyana*) and sect. *Drymosphace* (*S. miltiorrhiza*, *S. yunnanensis*, and *S. meiliensis*).

Most trees constructed from BI and ML analyses of each individual region produced a topology identical to that of the concatenated data set of 18 hypervariable regions; however, the *ccsA-ndhD* tree showed that *S. nilotica*, *S. sclarea*, *S. merjamie*, and two individuals of *S. officinalis* formed the first clade, and the second clade in turn was composed of *S. yangii*, *S. rosmarinus*, *S. splendens*, *S. hispanica*, and members of subg. *Glutinaria* (fig. S5*B*). Among the 18 trees, most clades were supported in the tree constructed by the single combined data set; the *ycf1* tree (fig. S5*R*) had the highest support values for most clades.

Discussion

This study is the first to examine plastome structure and variation among multiple subgenera of *Salvia*. We found that, in general, plastome structure is conserved within *Salvia*, although some taxa showed variation in terms of what gene/pseudogene occurred at the IR boundary. We also identified 626 short and

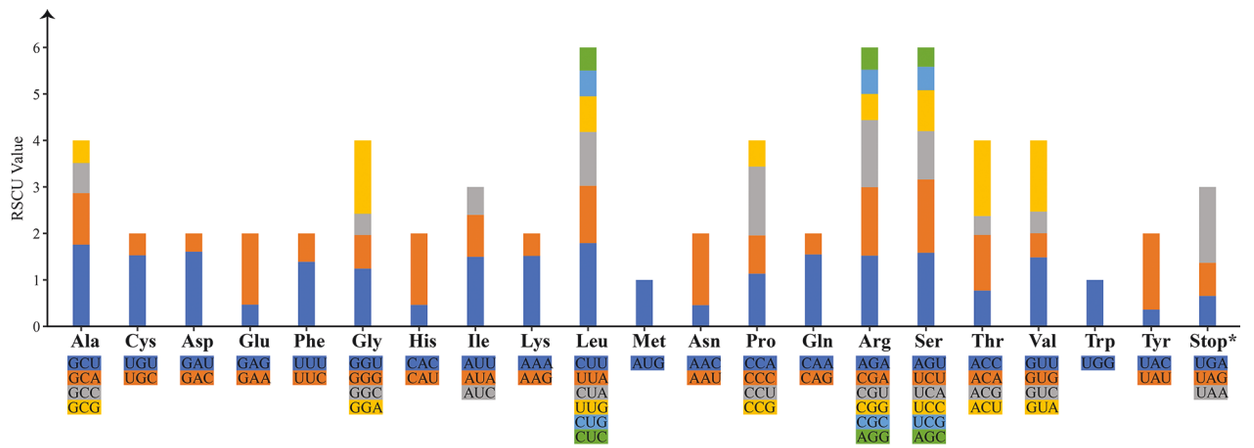

Fig. 7 Codon content of 20 amino acids and stop codon of 80 coding genes of the 17 *Salvia* plastomes. The color of the histogram corresponds to the color of the codons. RSCU = relative synonymous codon usage.

Table 7

Summary of the Data Set Information for Phylogenetic Analyses

Data matrix	Aligned (bp)	Constant sites (bp)	Variable sites (bp)	Parsimony informative (bp)
CPG	134,646	117,632 (87.36%)	10,417	6597 (4.90%)
CR	68,987	62,073 (89.98%)	4233	2681 (3.89%)
NCR	61,615	51,535 (83.64%)	6212	3868 (6.28%)
18 hypervariable regions	26,028	22,515 (86.50%)	3513	1774 (6.82%)

Note. CPG = complete plastid genome sequence; CR = coding region; NCR = noncoding region.

715 long repeat sequences among the *Salvia* plastomes. These repeats should prove valuable in future population genetics studies. Finally, we identified 18 hypervariable regions that should be useful in future phylogenetic studies as well as in bar coding efforts within *Salvia*. To our knowledge, this study is the first to identify SSRs and hypervariable regions within *Salvia*.

Comparison of the Plastomes in Salvia

All 17 *Salvia* plastomes share the same 114 unique genes (80 protein-coding genes, 30 tRNAs, and 4 rRNAs) that were recorded for *Origanum vulgare* L. from subfamily Nepetoideae (Lukas and Novak 2013). The gene order within *Salvia* plastomes is identical to that of other published Lamiaceae genomes (He et al. 2016; Jiang et al. 2017; Han and Zheng 2018), emphasizing the conserved nature of plastid genomes in seed plants, as suggested by Raubeson and Jansen (2005). Some protein-coding genes (e.g., *accD*, *ycf1*, *ycf2*, *rpl22*, *rps16*, *rpl23*, *infA*, and *ndhF*) have been shown to be independently lost over the course of angiosperm evolution (Millen et al. 2001; Jansen et al. 2007). However, these genes were not lost in any of the 17 *Salvia* plastomes. Furthermore, the sequence length and GC content of the whole plastomes, LSCs, SSCs, and IRs of the 17 *Salvia* plastomes were highly similar (table 3), suggesting that plastome structure and gene content are highly conserved in *Salvia* and other Lamiaceae.

The present study suggests that expansion and/or contraction of the IR regions is infrequent within *Salvia*, although IR expansion/contraction has occurred in many seed plants

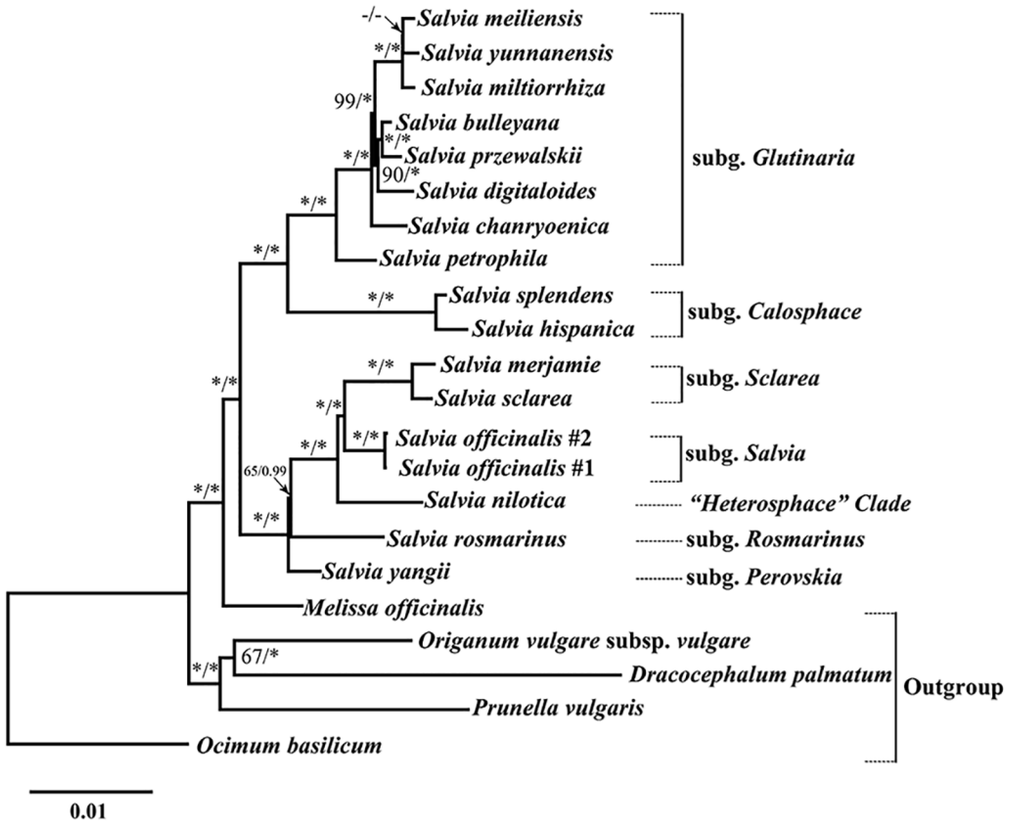

Fig. 8 Phylogenetic relationships of *Salvia* based on the data set of complete plastid genome sequences. The support values (bootstrap value [BS]/posterior probability [PP]) are indicated at the branches. BS and PP of 100% are indicated by an asterisk; BS values of <50% and PP support of <90% are indicated by a hyphen. The outgroup and recognized groups are marked in the right-hand bar.

(Cosner et al. 1997; Chumley et al. 2006; Liu et al. 2017; Zhou et al. 2018). Although we sampled only 16 species of *Salvia*, this sampling represented seven of the 10 *Salvia* subgenera. Among the 16 *Salvia* species, only slight IR contractions were observed in *S. hispanica* and *S. rosmarinus* (fig. 5). In *S. hispanica*, the IR contraction led to the IRa/LSC junction being located in the intergenic region between the *rps19* gene and the *rpl2* gene, while in *S. rosmarinus*, the contraction in the boundary of the IRa/SSC led to the *ndhF* gene being completely within the SSC region. In all other species, the IRb/LSC boundary was located between the *ψrps19* and *trnH*-GUG genes, and the junction of the IRb/SSC split the *ycf1* gene. There are probably additional IR expansions and contractions within *Salvia*, but this needs to be assessed by plastome analyses that include more comprehensive sampling.

Simple and Long Repeat Sequences

SSRs are known to be present in high numbers in angiosperm genomes (Yang et al. 2013) and are commonly considered to be the markers of choice for population genetics and evolutionary studies (Ebert and Peakall 2009). However, very few SSRs have been developed in *Salvia*. In this study, a total of 626 SSRs were detected among 16 *Salvia* species. As reported in other families (Wills et al. 2005; Ni et al. 2016), mononucleotide SSRs (especially of As or Ts) are the most abundant unit within *Salvia* plastomes and show meaningful differences among the species, from 16 (*S. yangii*) to 32 (*S. meiliensis* and *S. yunnanensis*; table 6). Other SSR units have no significant differences. In total, we found five unique repeat units (AAGG/CCTT, AATCT/AGATT, AAAATC /ATTTTG, AATTAT/AATTAT, and AAGTCT/ACTTAG) among the 16 *Salvia* species (table 6), which will contribute to further studies of the population genetics and phylogeography of the corresponding species.

With the exception of SSRs, repeats with lengths greater than 30 bp were considered to be long repeat sequences (Huang et al. 2019) in this study. Long repeat sequences are important because they are related to plastome organization (Salih et al. 2017) and may promote the rearrangement of the plastid genome (Timme et al. 2007; Cui et al. 2019; Huang et al. 2019) and increase the genetic diversity of populations (Qian et al. 2013). A total of 715 long repeat sequences was detected in the 17 *Salvia* plastomes. Compared with that of some angiosperm species (Liu et al. 2018; Menezes et al. 2018), the number of long repeat sequences in *Salvia* is quite high. For example, only 38 long repeat sequences were detected in *Scutellaria baicalensis* George. (Jiang et al. 2017). Among these repeats, the forward repeat was the most common, accounting for 48.95% of the total number of repeats, while the complementary repeats made up only 0.98%.

In some species of Lamiaceae (e.g., *S. baicalensis*), 80% of these repeats were found in intergenic regions (Jiang et al. 2017). In *Salvia* plastomes, however, more than half of all the repeats (56.6%) were located in CRs, mainly in the *accD*, *psaB*, *ycf1*, and *ycf2* genes, similar to other species within the lamiids (e.g., *Scrophularia dentata* Royle ex Benth.; Ni et al. 2016). Furthermore, forward and palindromic repeat sequences were common in all the species investigated in this study, while the reverse and complete repeat sequences were rare. All four types of long repeat sequences were detected only in *S. rosmarinus*. These unique long repeat sequences may have potential utility in the future study of this species.

Hypervariable Regions for Future Phylogenetic Studies within Salvia

Eighteen hypervariable regions were detected in the present study. Separate analyses conducted with single and combined data sets generated identical topologies, suggesting that these DNA regions are efficient molecular markers for phylogenetic study at low taxonomic levels, at least in *Salvia*.

Previously used plastid-based markers for *Salvia* phylogeny include *rbcL*, *trnL-trnF*, *psbA-trnH*, *ycf1*, *ycf1-rps15*, *rpl32-trnL*, *trnT-trnL*, *trnV* intron, *trnS-trnG*, *atpB-rbcL*, *rps16*, and *trnK-matK* (Walker et al. 2004, 2015; Taylor and Ayers 2006; Sudarmono and Okada 2007, 2008; Walker and Sytsma 2007; Jenks et al. 2011, 2013; Takano and Okada 2011; Drew and Sytsma 2012, 2013; Will and Claßen-Bockhoff 2014, 2017; Dizkirici et al. 2015; Will et al. 2015; Drew et al. 2017*a*, 2017*b*; Fragoso-Martínez et al. 2017, 2018; Hu et al. 2018). However, among these 12 regions, only three markers (*ycf1*, *rps15-ycf1*, and *rpl32-trnL*) were detected as hypervariable regions in the present study. Some plastome regions, such as the *accD*, *matK*, *rpoA*, *ycf2*, *ycf1*, and *rps7* genes, have high levels of divergence among families, and most of these sequences have been used for phylogenetic analyses (Huang et al. 2010; Domenech et al. 2014; Luo et al. 2014; Bodin et al. 2016). Among those DNA regions, however, only *ycf1* and *matK* have been shown to be useful for inferring phylogenetic relationships within *Salvia* species (e.g., Walker and Sytsma 2007; Drew and Sytsma 2011, 2012).

Although it has not been widely used because of a lack of universal primers and the large number of primer pairs needed to sequence the entire region, the *ycf1* gene is a useful marker for phylogenetic studies. Neubig et al. (2008) demonstrated that *ycf1* is the most rapidly evolving gene in orchids and is useful for phylogenetic study. Later, Drew and Sytsma (2011) suggested that *ycf1* has great phylogenetic utility at different taxonomic levels in Lamiaceae, from closely related species to between subfamilies. Drew and Sytsma (2011) showed that *ycf1* was 50% more informative than *trnL-F* within *Lepechinia* and other genera from tribe Mentheae (Lamiaceae). In our analyses, informative sites from *ycf1* (9.4%) are three times higher than those from *trnL-F* (less than 3%) and generated the phylogenetic tree with the highest support values (fig. S5). Drew and Sytsma (2011) elaborated on *ycf1* structure, placement, and evolution, including the placement of *ycf1* at the intersection of the IR and SSC regions; the expansion or contraction of the IR, which can cause *ycf1* to become embedded within the IR; and the independent loss in some land plant plastid genomes. Drew and Sytsma (2011) cautioned that *ycf1* could be problematic in phylogenetic analyses because it has essentially been lost in some angiosperm lineages (e.g., Poaceae), but in our analyses of *Salvia*, the topology of the phylogenetic tree based on only the *ycf1* gene was identical to that of the combined data set (fig. S4) and very similar to the tree topology produced by most individual regions (fig. S5*H*–S5*L*). Thus, we suggest that *ycf1* will be an excellent marker for phylogenetic inferences in future studies within *Salvia*.

Some NCRs, such as *rpl32-trnL-ccsA* (9.12%), *rps15-ycf1* (8.64%), *petA-psbJ* (7.11%), and *psbK-psbI* (7.06%), also contained a high number of informative sites, which indicates

that they may also be useful for phylogenetic studies of *Salvia*. Of these regions, *rps15-ycf1* and *rpl32-trnL* have previously been used in systematic studies of *Salvia* (Drew and Sytsma 2012, 2013; Will and Claßen-Bockhoff 2014, 2017; Walker et al. 2015; Will et al. 2015; Hu et al. 2018). The *petA-psbJ* and *psbK-psbI* regions were used for phylogenetic reconstruction and/or biogeographic studies within the Lamiaceae (Gao et al. 2014; Welch et al. 2016; Zhang et al. 2017*a*) but have not been widely used.

Phylogenetic Inference

We used four data sets (complete plastome, protein-coding exons, NCRs, and combined 18 hypervariable regions) to reconstruct the phylogeny of *Salvia*. In all analyses, the monophyly of *Salvia* was recovered with high support (figs. 8, S2–S4). Topologies were generally congruent; however, the position of *S. rosmarinus* and *S. yangii* varied. In the trees generated from the NCR data set (fig. S3) and the combined 18 hypervariable regions (fig. S4), *S. yangii* and *S. rosmarinus* formed a sister clade to *S. nilotica*, *S. officinalis*, *S. sclarea*, and *S. merjamie*, but support values were low. In the trees of the CR (fig. S2) and CPG data sets (fig. 8), *S. yangii* is the first-diverging lineage, followed by *S. rosmarinus*, *S. nilotica*, *S. officinalis*, *S. sclarea*, and *S. splendens*. By using nrITS and nrETS sequences, Hu et al. (2018) also recovered this topology and likewise reported low support values for it. Recently, these relationships were confirmed by Kriebel et al. (2019) using nuclear loci obtained by anchored hybrid enrichment (Lemmon et al. 2012).

Earlier studies revealed that both the phylogenetic resolution and the support values of branches may be considerably improved by more and longer DNA sequences (Rokas and Carroll 2005; Philippe et al. 2011), and our analyses reached a similar conclusion. In this study, phylogenetic trees based on combined data sets of protein-coding regions (CR; 68,987 bp) and complete whole genome (CPG; 134,646 bp) sequences generated higher support values than the other two smaller data sets (i.e., NCR and the combined data set of 18 hypervariable regions), even though the percentages of parsimony-informative sites (3.89% and 4.90%, respectively; table 7) of these two data sets were lower than those of the NCR (6.28%) and the combined 18 hypervariable regions (6.82%). This result is to be expected because the phylogenetic methods used here are not based on the number of parsimony-informative sites; thus, this metric is an underestimation of the relevant sites for phylogenetic reconstruction. Furthermore, our analysis found that NCRs across *Salvia* plastomes possessed the highest sequence divergence among the four data sets tested (table 7), while the phylogeny based on NCRs did not have the highest branch support. The relatively lower branch supports found in the tree (fig. S3) may be attributed to the faster (and sometimes aleatory) mutation rates of NCRs in the plastome, as Wiens (2003) has suggested.

Our results also support the monophyly of *Salvia* subg. *Glutinaria* (figs. 8, S2–S4), which was recently established by Hu et al. (2018) and consists of nearly all the native *Salvia* found in East Asia (Hu et al. 2020). These authors also recognized eight sections within the subgenus. Here, we included eight species representing four sections (i.e., *Sonchifoliae*, *Glutinaria*, *Eurysphace*, and *Drymosphace*) of this subgenus for analysis. Although only four sections were represented by eight species in the present study, the topology is comparable to that of Hu et al. (2018), with *S. petrophila* as the first-diverging species within this subgenus. Within subg. *Glutinaria*, relationships among three species (*S. meiliensis*, *S. yunnanensis*, and *S. miltiorrhiza*) within sect. *Drymosphace* varied in different data sets. Interspecific relationships within this section were unresolved in Hu et al. (2018). In the present study, the plastome phylogenies corroborate several previously reconstructed relationships within *Salvia* (Hu et al. 2018; Kriebel et al. 2019), indicating that the complete plastome sequences and 18 hypervariable regions provide a well-resolved and well-supported tree for the backbone relationships of *Salvia*. However, resolving relationships at the species level is still challenging.

In this study, 17 plastid genomes of *Salvia* were analyzed. We evaluated the efficacy of complete plastomes for phylogenetic reconstruction within *Salvia*, focusing on subg. *Glutinaria*. The use of whole plastomes produced well-supported and highly resolved phylogenies in this phylogenetically recalcitrant group, demonstrating the utility of plastomes for phylogenetic reconstruction within *Salvia*. In addition, 18 hypervariable regions were identified here; these regions display much more variation than most markers previously used for phylogenetic studies in *Salvia* and have potential as novel plastid markers for phylogenetic, species discrimination (DNA bar coding), and population genetics studies. This case study suggests that complete plastomes can substantially increase the resolution of phylogenetic trees and will provide new insights into speciation and lineage diversification in both *Salvia* and other groups of mints.

Acknowledgments

We thank Dr. Shi-Jin Li for his assistance in sample collection of *Salvia merjamie* and *S. nilotica*. This study was supported by the Excellent Young Scholars Yunnan program (2019FI009), the Ten Thousand Talents of Yunnan program (YNWR-QNBJ-2018-279), and the CAS Light of West China program's support of Chun-Lei Xiang. B. Drew acknowledges the US National Science Foundation for funding (DEB-1655611), and G.-X. Hu acknowledges the Natural Science Foundation of Guizhou Province (20161049).

Literature Cited

Aronesty E 2013 Comparison of sequencing utility programs. Open Bioinformat J 7:1–8. https://doi.org/10.2174/1875036201307010001.

Barrett CF, CD Specht, J Leebens-Mack, DW Stevenson, WB Zomlefer, JI Davis 2014 Resolving ancient radiations: can complete plastid gene sets elucidate deep relationships among the tropical gingers (Zingiberales)? Ann Bot 113:119–133.

Bodin SS, JS Kim, JH Kim 2016 Phylogenetic inferences and the evolution of plastid DNA in Campynemataceae and the mycoheterotrophic *Corsia dispar* D. L. Jones & B. Gray (Corsiaceae). Plant Mol Biol Report 34:192–210.

Chen C, W Hua 2019 The complete chloroplast genome of rosemary (*Rosmarinus officinalis*). Mitochondrial DNA B 4:147–148.

Chen H, J Zhang, G Yuan, C Liu 2014 Complex interplay among DNA modification, noncoding RNA expression and protein-coding RNA expression in *Salvia miltiorrhiza* chloroplast genome. PLoS ONE 9:e99314.

Chen YP, TC Wilson, YD Zhou, ZH Wang, ED Liu, H Peng, CL Xiang 2019 *Isodon hsiwenii* (Lamiaceae: Nepetoideae), a new species from Yunnan, China. Syst Bot 44:913–922.

Chumley TW, JD Palmer, JP Mower, HM Fourcade, PJ Calie, JL Boore, RK Jansen 2006 The complete chloroplast genome sequence of *Pelargonium* × *hortorum*: organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. Mol Biol Evol 23:2175–2190.

Claßen-Bockhoff R, P Wester, E Tweraser 2003 The staminal lever mechanism in *Salvia* L. (Lamiaceae): a review. Plant Biol 5:33–41.

Cosner ME, RK Jansen, JD Palmer, SR Downie 1997 The highly rearranged chloroplast genome of *Trachelium caeruleum* (Campanulaceae): multiple inversions, inverted repeat expansion and contraction, transposition, insertions/deletions, and several repeat families. Curr Genet 31:419–429.

Couvreur TLP, AJ Helmstetter, EJM Koenen, K Bethune, RD Brandão, SA Little, H Sauquet, RHJ Erkens 2019 Phylogenomics of the major tropical plant family Annonaceae using targeted enrichment of nuclear genes. Front Plant Sci 9:1941.

Cui YX, XL Chen, LP Nie, W Sun, HY Hu, YL Lin, HT Li, XL Zheng, JY Song, H Yao 2019 Comparison and phylogenetic analysis of chloroplast genomes of three medicinal and edible *Amomum* species. Int J Mol Sci 20:4040.

Darling AC, B Mau, FR Blattner, NT Perna 2004 Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res 14:1394–1403.

Ding BY, ZH Chen, YL Xu, XF Jin, DF Wu, JB Chen, WJ Wu 2019 New species and combination of Lamiaceae from Zhejiang, China. Guihaia 39:10–15.

Dizkirici A, F Celep, C Kansu, A Kahraman, M Dogan, Z Kaya 2015 A molecular phylogeny of *Salvia euphratica* sensu lato (*Salvia* L., Lamiaceae) and its closely related species with a focus on the section *Hymenosphace*. Plant Syst Evol 301:2313–2323.

Domenech B, CB Asmussen-Lange, WJ Baker, E Alapetite, JC Pintaud, S Nadot 2014 A phylogenetic analysis of palm subtribe Archontophoenicinae (Arecaceae) based on 14 DNA regions. Bot J Linn Soc 175:469–481.

Doyle JJ, JL Doyle 1987 A rapid DNA isolation procedure for small amounts of fresh leaf tissue. Phytochem Bull 19:11−15.

Drew BT, JG González-Gallegos, CL Xiang, R Kriebel, CP Drummond, JB Walker, KJ Sytsma 2017*a* *Salvia* united: the greatest good for the greatest number. Taxon 66:133–145.

Drew BT, R Kriebel, L Kahan, JG González-Gallegos, F Celep, ER Lemmon, AR Lemmon, KJ Sytsma 2020 Anchored hybrid enrichment yields unprecedented insights into the phylogenetics of *Salvia*. Int J Plant Sci 181:XXX–XXX.

Drew BT, S Liu, JM Bonifacino, KJ Sytsma 2017*b* Amphitropical disjunctions in New World Menthinae: three Pliocene dispersals to South America following late Miocene dispersal to North America from the Old World. Am J Bot 104:1695–1707.

Drew BT, KJ Sytsma 2011 Testing the monophyly and placement of *Lepechinia* in the tribe Mentheae (Lamiaceae). Syst Bot 36:1038–1049.

——— 2012 Phylogenetics, biogeography, and staminal evolution in the tribe Mentheae (Lamiaceae). Am J Bot 99:933–953.

——— 2013 The South American radiation of *Lepechinia* (Lamiaceae): phylogenetics, divergence times and evolution of dioecy. Bot J Linn Soc 171:171–190.

Du Y, YY Wang, CL Xiang, MQ Yang 2019 Characterization of the complete chloroplast genome of *Salvia przewalskii* Maxim. (Lamiaceae), a substitute for Dan-Shen *Salvia miltiorrhiza* Bunge. Mitochondrial DNA B 4:981–982.

Ebert D, R Peakall 2009 Chloroplast simple sequence repeats (cpSSRs): technical resources and recommendations for expanding cpSSR discovery and applications to a wide array of plant species. Mol Ecol Resour 9:673–690.

Fragoso-Martínez I, M Martínez-Gordillo, GA Salazar, F Sazatornil, AA Jenks, M del Rosario García Peña, G Barrera-Aveleida, et al 2018 Phylogeny of the Neotropical sages (*Salvia* subg. *Calosphace*; Lamiaceae) and insights into pollinator and area shifts. Plant Syst Evol 304:43–55.

Fragoso-Martínez I, GA Salazar, M Martínez-Gordillo, S Magallón, L Sánchez-Reyes, EM Lemmon, AR Lemmon, F Sazatornil, CG Mendoza 2017 A pilot study applying the plant anchored hybrid enrichment method to New World sages (*Salvia* subgenus *Calosphace*; Lamiaceae). Mol Phylogenet Evol 117:124–134.

Frazer KA, L Pachter, A Poliakov, EM Rubin, I Dubchak 2004 VISTA: computational tools for comparative genomics. Nucleic Acids Res 32:W273–W279.

Gao XY, HH Meng, ML Zhang 2014 Diversification and vicariance of desert plants: evidence inferred from chloroplast DNA sequence variation of *Lagochilus ilicifolius* (Lamiaceae). Biochem Syst Ecol 55:93–100.

González-Gallegos JG 2015 Two new *Salvia* species (Lamiaceae) from the Sierra Madre Occidental, Durango, Mexico. Syst Bot 40:1093–1101.

González-Gallegos JG, BY Bedolla-García, G Cornejo-Tenorio, JL Fernández-Alonso, I Fragoso-Martínez, MR García-Peña, RM Harley, et al 2020 Richness and distribution of *Salvia* subgenus *Calosphace* (Lamiaceae). Int J Plant Sci 181:XXX–XXX.

González-Gallegos JG, OJ Gama-Villanueva 2013 Resurrection of *Salvia* species (Lamiaceae) recently synonymized in Flora Mesoamericana. Phytotaxa 151:1–24.

Ha YH, KS Choi, K Choi 2018 Characterization of complete chloroplast genome of endemic species of Korea Peninsular, *Salvia chanryoenica* (Lamiaceae). Mitochondrial DNA B 3:992–993.

Han YW, TY Zheng 2018 The complete chloroplast genome of the common self-heal, *Prunella vulgaris* (Lamiaceae). Mitochondrial DNA B 4:147–148.

He Y, H Xiao, C Deng, L Xiong, J Yang, C Peng 2016 The complete chloroplast genome sequences of the medicinal plant *Pogostemon cablin*. Int J Mol Sci 17:820.

He YH, LM Han, YP Liu, N Tian, X Su, ZZ Wang 2017 Complete sequence analysis of chloroplast genome of *Salvia japonica*. Bull Bot Res 37:572–578.

Hu GX, ED Liu, ZK Wu, KJ Sytsma, BT Drew, CL Xiang 2020 Integrating DNA sequences with morphological analysis clarifies phylogenetic position of *Salvia grandifolia* (Lamiaceae): an enigmatic species endemic to southwestern China. Int J Plant Sci 181:XXX–XXX.

Hu GX, Y Liu, WB Xu, ED Liu 2014 *Salvia petrophila* sp. nov. (Lamiaceae) from north Guangxi and south Guizhou, China. Nord J Bot 32:190–195.

Hu GX, ED Liu, T Zhang, J Cai, CL Xiang 2017 *Salvia luteistriata* (Lamiaceae), a new species from northeastern Sichuan, China. Phytotaxa 314:123–128.

Hu GX, A Takano, BT Drew, ED Liu, DE Soltis, PS Soltis, H Peng, CL Xiang 2018 Phylogeny and staminal evolution of *Salvia* (Lamiaceae, Nepetoideae) in East Asia. Ann Bot 122:649–668.

Huang JL, GL Sung, DM Zhang 2010 Molecular evolution and phylogeny of the angiosperm *ycf2* gene. J Syst Evol 48:240–248.

Huang YY, ZR Yang, S Huang, WL An, J Li, XS Zheng 2019 Comprehensive analysis of *Rhodomyrtus tomentosa* chloroplast genome. Plants 8:89.

Jansen RK, Z Cai, LA Raubeson, H Daniell, CW de Paphilis, J Leebens-Mack, KF Müller, et al 2007 Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci USA 104:19369–19374.

Jansen RK, TA Ruhlman 2012 Plastid genomes of seed plants. Pages 103–126 *in* R Bock, V Knoop, eds. Genomics of chloroplasts and mitochondria. Springer, Dordrecht.

Jansen RK, C Saski, SB Lee, AK Hansen, H Daniell 2011 Complete plastid genome sequences of three rosids (*Castanea*, *Prunus*, *Theobroma*): evidence for at least two independent transfers of *rpl22* to the nucleus. Mol Biol Evol 28:835–847.

Jenks AA, JB Walker, SC Kim 2011 Evolution and origins of the Mazatec hallucinogenic sage, *Salvia divinorum* (Lamiaceae): a molecular phylogenetic approach. J Plant Res 124:593–600.

——— 2013 Phylogeny of New World *Salvia* subgenus *Calosphace* (Lamiaceae) based on cpDNA (*psbA-trnH*) and nrDNA (ITS) sequence data. J Plant Res 126:483–496.

Jiang D, Z Zhao, T Zhang, W Zhong, C Liu, Q Yuang, L Huang 2017 The chloroplast genome sequence of *Scutellaria baicalensis* provides insight into intraspecific and interspecific chloroplast genome diversity in *Scutellaria*. Genes 8:227.

Jin JJ, WB Yu, JB Yang, Y Song, TS Yi, DZ Li 2018 GetOrganelle: a simple and fast pipeline for de novo assembly of a complete circular chloroplast genome using genome skimming data. bioRxiv 256479.

Katoh K, DM Standley 2013 MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780.

Kearse M, R Moir, A Wilson, S Stones-Havas, M Cheung, S Sturrock, S Buxton, et al 2012 Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647–1649.

Kriebel R, BT Drew, CP Drummond, JG González-Gallegos, F Celep, MM Mahdjoub, JP Rose, et al 2019 Tracking temporal shifts in area, biomes, and pollinators in the radiation of *Salvia* (sages) across continents: leveraging anchored hybrid enrichment and targeted sequence data. Am J Bot 106:573–597.

Kurtz S, JV Choudhuri, E Ohlebusch, C Schleiermacher, J Stoye, R Giegerich 2001 REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res 29:4633–4642.

Langmead B, SL Salzberg 2012 Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359.

Lemmon AR, SA Emme, EM Lemmon 2012 Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst Biol 61:727–744.

Li HW, IC Hedge 1994 *Salvia*. Pages 196–224 *in* CY Wu, PH Raven, DY Hong, eds. Flora of China. Vol 17. Science, Beijing/Missouri Botanical Gardens, St. Louis.

Li QQ, MH Li, QJ Yuan, ZH Cui, LQ Huang, PG Xiao 2013 Phylogenetic relationships of *Salvia* (Lamiaceae) in China: evidence from DNA sequence datasets. J Syst Evol 51:184–195.

Liang CL, L Wang, J Lei, BZ Duan, WS Ma, SM Xiao, HJ Qi, et al 2019 A comparative analysis of the chloroplast genomes of four *Salvia* medicinal plants. Engineering 5:907–915.

Lin CP, JP Huang, CS Wu, CY Hsu, SM Chaw 2010 Comparative chloroplast genomics reveals the evolution of Pinaceae genera and subfamilies. Genome Biol Evol 2:504–517.

Liu H, J He, C Ding, R Lyu, L Pei, J Cheng, L Xie 2018 Comparative analysis of complete chloroplast genomes of *Anemoclema*, *Anemone*, *Pulsatilla*, and *Hepatica* revealing structural variations among genera in tribe Anemoneae (Ranunculaceae). Front Plant Sci 9:1097.

Liu LX, R Li, JRP Worth, X Li, P Li, KM Cameron, CX Fu 2017 The complete chloroplast genome of Chinese bayberry (*Morella rubra*, Myricaceae): implications for understanding the evolution of Fagales. Front Plant Sci 8:968.

Lohse M, O Drechsel, S Kahlau, R Bock 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 TM, PP Chan 2016 tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. Nucleic Acids Res 44:W54–W57.

Lukas B, J Novak 2013 The complete chloroplast genome of *Origanum vulgare* L. (Lamiaceae). Gene 528:163–169.

Luo J, BW Hou, ZT Niu, W Liu, QY Xue, XY Ding 2014 Comparative chloroplast genomes of photosynthetic orchids: insights into evolution of the Orchidaceae and development of molecular markers for phylogenetic applications. PLoS ONE 9:e99016.

Ma PF, YX Zhang, CX Zeng, ZH Guo, DZ Li 2014 Chloroplast phylogenomic analyses resolve deep-level relationships of an intractable bamboo tribe Arundinarieae (Poaceae). Syst Biol 64:933–950.

Menezes APA, LC Resende-Moreira, RSO Buzatti, AG Nazareno, M Carlsen, FP Lobo, E Kalapothakis, MB Lovata 2018 Chloroplast genomes of *Byrsonima* species (Malpighiaceae): comparative analysis and screening of high divergence sequences. Sci Rep 8:2210.

Millen RS, RG Olmstead, KL Adams, JD Palmer, NT Lao, L Heggie, TA Kavanagh, et al 2001 Many parallel losses of *infA* from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. Plant Cell 13:645–658.

Miller MA, P Wayne, T Schwartz 2010 Creating the CIPRES Science Gateway for inference of large phylogenetic trees. Paper presented at the 2010 Gateway Computing Environments Workshop, New Orleans, November 14.

Moore MJ, CD Bell, PS Soltis, DE Soltis 2007 Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. Proc Natl Acad Sci USA 104:19363–19368.

Moore MJ, PS Soltis, CD Bell, JG Burleigh, DE Soltis 2010 Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. Proc Natl Acad Sci USA 107:4623–4628.

Neubig KM, WM Whitten, BS Carlsward, MA Blanco, L Endara, NH Williams, M Moore 2008 Phylogenetic utility of *ycf1* in orchids: a plastid gene more variable than *matK*. Plant Syst Evol 277:75–84.

Ni L, Z Zhao, G Dorje, L Ma 2016 The complete chloroplast genome of Ye-Xing-Ba (*Scrophularia dentata*; Scrophulariaceae), an alpine Tibetan herb. PLoS ONE 11:e0158488.

Peng H, CL Xiang 2017 Lamiaceae. Pages 1–592 *in* H Peng, ed. Medicinal flora of China. Vol 9. Peking University Medical Press, Beijing.

Philippe H, H Brinkmann, DV Lavrov, DTJ Littlewood, M Manuel, G Wörheide, D Baurain 2011 Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol 9:e1000602.

Powell W, M Morgante, C Andre, M Hanafey, J Vogel, S Tingey, A Rafalsk 1996 The comparison of RFLP, RAPD, AFLP and SSRP (microsatellite) markers for germplasm analysis. Mol Breed 2:225–238.

Qian J, J Song, H Gao, Y Zhu, J Xu, X Pang, H Yao, et al 2013 The complete chloroplast genome sequence of the medicinal plant *Salvia miltiorrhiza*. PLoS ONE 8:e57607.

Rabah SO, C Lee, NH Hajrah, RM Makki, HF Alharby, AM Alhebshi, JSM Sabir, RK Jansen, TA Ruhlman 2017 Plastome sequencing of ten nonmodel crop species uncovers a large insertion of mitochondrial DNA in cashew. Plant Genome 10:3835.

Rambaut A, MA Suchard, AJ Drummond 2014 Tracer v1.6. http://beast.bio.ed.ac.uk/Tracer.

Raubeson LA, RK Jansen 2005 Chloroplast genome of plants. Pages 45–68 *in* RJ Henry, ed. Plant diversity and evolution: genotypic and phenotypic variation in higher plants. CABI, Oxfordshire, UK.

Rokas A, SB Carroll 2005 More genes or more taxa? the relative contribution of gene number and taxon number to phylogenetic accuracy. Mol Biol Evol 22:1337–1344.

Ronquist F, M Teslenko, P Van der Mark, DL Ayres, A Darling, S Höhna, B Larget, L Liu, MA Suchard, JP Huelsenbeck 2012 MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61:539–542.

Rozas J, A Ferrer-Mata, JC Sánchez-Delbarrio, S Guirao-Rico, P Librado, SE Ramos-Onsins, A Sánchez-Gracia 2017 DnaSP 6: DNA sequence polymorphism analysis of large data sets. Mol Biol Evol 34:3299–3302.

Salih RHM, L Majeský, T Schwarzacher, R Gornall, P Heslop-Harrison 2017 Complete chloroplast genomes from apomictic *Taraxacum* (Asteraceae): identity and variation between three microspecies. PLoS ONE 12:e0168008.

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

Stöver B, K Müller 2010 TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses. BMC Bioinformatics 11:1–9.

Stull GW, RD de Stefano, DE Soltis, PS Soltis 2015 Resolving basal lamiid phylogeny and the circumscription of Icacinaceae with a plastome-scale data set. Am J Bot 102:1794–1813.

Sudarmono, H Okada 2007 Speciation process of *Salvia isensis* (Lamiaceae), a species endemic to serpentine areas in the Ise-Tokai district, Japan, from the viewpoint of the contradictory phylogenetic trees generated from chloroplast and nuclear DNA. J Plant Res 120:483–490.

——— 2008 Genetic differentiations among the populations of *Salvia japonica* (Lamiaceae) and its related species. Hayati J Biosci 15:18–26.

Takano A, H Okada 2011 Phylogenetic relationships among subgenera, species, and varieties of Japanese *Salvia* L. (Lamiaceae). J Plant Res 124:245–252.

Taylor RM, TJ Ayers 2006 Systematics of *Salvia pachyphylla* (Lamiaceae). Madrono 53:11–24.

Thiel T, W Michalek, R Varshney, A Graner 2003 Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theor Appl Genet 106:411–422.

Timme RE, JV Kuehl, JL Boore, RK Jansen 2007 A comparative analysis of the *Lactuca* and *Helianthus* (Asteraceae) plastid genomes: identification of divergent regions and categorization of shared repeats. Am J Bot 94:302–312.

Walker JB, BT Drew, KJ Sytsma 2015 Unravelling species relationships and diversification within the iconic California Floristic Province sages (*Salvia* subgenus *Audibertia*, Lamiaceae). Syst Bot 40:826–844.

Walker JB, KJ Sytsma 2007 Staminal evolution in the genus *Salvia* (Lamiaceae): molecular phylogenetic evidence for multiple origins of the staminal lever. Ann Bot 100:375–391.

Walker JB, KJ Sytsma, J Treutlein, M Wink 2004 *Salvia* (Lamiaceae) is not monophyletic: implications for the systematics, radiation, and ecological specializations of *Salvia* and tribe Mentheae. Am J Bot 91:1115–1125.

Welch AJ, K Collins, A Ratan, D Drautz-Moses, SC Schuster, C Lindqvist 2016 The quest to resolve recent radiations: plastid phylogenomics of extinct and endangered Hawaiian endemic mints (Lamiaceae). Mol Phylogenet Evol 99:16–33.

Wick RR, MB Schultz, J Zobel, KE Holt 2015 Bandage: interactive visualization of de novo genome assemblies. Bioinformatics 31:3350–3352.

Wickham H 2009 ggplot2: elegant graphics for data analysis. Springer, New York.

Wiens JJ 2003 Missing data, incomplete taxa, and phylogenetic accuracy. Syst Biol 52:528–538.

Will M, R Claßen-Bockhoff 2014 Why Africa matters: evolution of Old World *Salvia* (Lamiaceae) in Africa. Ann Bot 114:61–83.

——— 2017 Time to split *Salvia* s.l. (Lamiaceae): new insights from Old World *Salvia* phylogeny. Mol Phylogenet Evol 109:33–58.

Will M, N Schmalz, R Claßen-Bockhoff 2015 Towards a new classification of *Salvia* s.l.: (re)establishing the genus *Pleudia* Raf. Turk J Bot 39:693–707.

Williams AV, JT Miller, I Small, PG Nevill, LM Boykin 2016 Integration of complete chloroplast genome sequences with small amplicon datasets improves phylogenetic resolution in *Acacia*. Mol Phylogenet Evol 96:1–8.

Wills DM, ML Hester, AZ Liu, JM Burke 2005 Chloroplast SSR polymorphisms in the Compositae and the mode of organellar inheritance in *Helianthus annuus*. Theor Appl Genet 110:941–947.

Wyman SK, RK Jansen, JL Boore 2004 Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20:3252–3255.

Xi Z, BR Ruhfel, H Schaefer, AM Amorim, M Sugumaran, KJ Wurdack, PK Endress, et al 2012 Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. Proc Natl Acad Sci USA 109:17519–17524.

Xia X, Z Xie 2001 DAMBE: software package for data analysis in molecular biology and evolution. J Hered 92:371–373.

Xiang CL, HJ Dong, S Landrein, F Zhao, WB Yu, DE Soltis, PS Soltis, et al 2020 Revisiting the phylogeny of Dipsacales: new insights from phylogenomic analyses of complete plastome sequences. J Syst Evol 58:103–117.

Xiang CL, GX Hu, H Peng 2016 *Salvia wuana* (Lamiaceae), a new name for *S. pauciflora* E. Peter. Phytotaxa 255:99–100.

Yang JB, SX Yang, HT Li, J Yang, DZ Li 2013 Comparative chloroplast genomes of *Camellia* species. PLoS ONE 8:e73053.

Yu XQ, LM Gao, DE Soltis, PS Soltis, JB Yang, L Fang, SX Yang, DZ Li 2017 Insights into the historical assembly of East Asian subtropical evergreen broadleaved forests revealed by the temporal history of the tea family. New Phytol 215:1235–1248.

Zeng L, Q Zhang, R Sun, N Zhang, H Ma 2014 Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. Nat Commun 5:4956.

Zhang LN, PF Ma, YX Zhang, CX Zeng, L Zhao, DZ Li 2019 Using nuclear loci and allelic variation to disentangle the phylogeny of *Phyllostachys* (Poaceae, Bambusoideae). Mol Phylogenet Evol 137:222–235.

Zhang ML, XQ Zeng, SC Sanderson, VV Byalt, AP Sukhorukov 2017*a* Insight into central Asian flora from the Cenozoic Tianshan montane origin and radiation of *Lagochilus* (Lamiaceae). PLoS ONE 12:e0178389.

Zhang SD, JJ Jin, SY Chen, MW Chase, DE Soltis, HT Li, JB Yang, DZ Li, TS Yi 2017*b* Diversification of Rosaceae since the late Cretaceous based on plastid phylogenomics. New Phytol 214:1355–1367.

Zhong B, T Yonezawa, Y Zhong, M Hasegawa 2010 The position of Gnetales among seed plants: overcoming pitfalls of chloroplast phylogenomics. Mol Biol Evol 27:2855–2863.

Zhou T, J Wang, Y Jia 2018 Comparative chloroplast genome analyses of species in *Gentiana* section *Cruciata* (Gentianacea) and the development of authentication markers. Int J Mol Sci 19:1962.