Comparative genomics and phylogenetic analysis of the chloroplast genomes in three medicinal Salvia species for bioexploration

Qing Du
Qinghai Minzu University
Jing ZENG
Chinese Academy of Medical Sciences, Peking Union Medical College
Liqiang Wang
Heze University
Zhuoer Chen
Xiangnan University
Junchen Zhou
Xiangnan University
SIHUI Sun
Xiangnan University
BIN WANG
Xiangnan University
CHANG LIU (cliu6688@yahoo.com)
Chinese Academy of Medical Sciences, Peking Union Medical College

Research Article

Keywords: Salvia bowleyana, Salvia splendens, Salvia officinalis, Chloroplast genome, Comparative genomics, Repeat analysis, Hypervariable regions, DNA barcode, Phylogenetic analysis

Posted Date: May 11th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1582501/v1

Abstract

Background: *Salvia bowleyana, S. splendens*, and *S. officinalis* are globally distributed and have been widely used to treat coronary heart disease, liver tumors, and viral diseases. To systematically determine their phylogenetic relationship and develop molecular markers for species determination, we sequenced and assembled their chloroplast genomes, and analyzed the genome characteristics. Moreover, we compared the phylogenetic specification and divergence genes fragments of chloroplast genomes from the three *Salvia* species.

Results: The length of *S. bowleyana*, *S. splendens*, and *S. officinalis* chloroplast genomes were 151387 bp, 150604 bp, and 151163 bp, respectively. The sizes of the large-single copy, small-single copy, and inverted repeat regions were 82772 bp, 17573 bp, and 51042 bp for *S. bowleyana*; 82181 bp, 17857 bp, and 50566 bp for *S. splendens*; 82429 bp, 17510 bp, and 51224 bp for *S. officinalis*, respectively. The GC contents of the three chloroplast genomes were 38.01%, 38.04%, 38.04%, partly. In the comparison of chloroplast genomes from Lamiaceae family, the six genes *ndh*B, *rpl*2, *rpl*23, *rps*7, *rps*12, and *ycf*2 were present in the IRs regions of all 41 chloroplast genomes in the Lamiaceae family. We found that one of the gene *rpl*20 was intact and stably occurred in all 41 species chloroplast genomes, however, another pseudogene one was lost in that of 40 species except the *Dracocephalum heterophyllum*. For the repeat analysis, 29 tandem repeats, 35, 29, 24 simple-sequence repeats(SSRs), and 47, 49, 40 interspersed repeats were identified in the three *Salvia* species chloroplast genomes based on the diverse requirements. The three specific intergenic sequences(IGS) of *rps*16-*trn*Q-UUG, *trn*L-UAA-*trn*F-GAA, and *trn*M-CAU-*atp*E were found to discriminate the certain species by comparing 23 *Salvia* chloroplast genomes. Six genes including *rpl*22, *rps*19, *rpl*2, *ycf*1, *ndh*F, and *psb*A were found in the highly diverse IR boundary regions. The genetic distance analysis of IGS showed the *trn*L-UAG-*ccs*A, *rps*16-*trn*Q-UUG, *ccs*A-*ndh*D, *rps*15-*ycf*1, and *ndh*E-*ndh*G regions had the higher variability. Furthermore, the phylogenetic tree inferred that the 23 *Salvia* species formed a monophyletic group. Lastly, two pairs of Genus-specific DNA barcode primers were identified, which can be used to amplify the part sequence of *trn*M-CAU-*atp*E and *ccs*A-*ndh*D region.

Conclusions: We acquired the complete chloroplast genome of the three *Salvia* species, which will provide a solid foundation to understand their phylogenetic status in *Salvia* genus. Moreover, the research can provide the probability to discriminate the *Salvia* species compared with the genomics between the phenotype and the distinction of gene fragments .

Background

The Lamiaceae is a large family, including 10 subfamilies, 220 genera, and 3,500 species mainly distributed in the area of Asia, Africa, and Europe. In historical evolutionary, the family of Lamiaceae is most closely related to the the family of Verbenaceae and Violinaceae [1, 2]. In China, 99 genera and more than 800 species in the Lamiaceae family are found, which contain 1050 *Salvia* species. Among them, 78 varied species and 32 variants mostly grow in the tropical or temperate areas [3]. The species from the Lamiaceae family are famous for containing a variety of aromatic oils, many of which are available for medicinal applications. The most well known species include *Scutellaria baicalensis*, *Salvia miltiorrhiza, Agastache rugosa, Leonurus japonicus, Mentha canadensis, Nepeta cataria, Perilla frutescens, Elsholtzia ciliata, Thymus mongolicus, Lavandula angustifolia, Rosmarinus officinalis*, etc. Their active ingredients have the diverse activities such as insecticidal, antibacterial, or weeding [4] and can be develop into the products of plant-derived pesticides [5].

The medicinal part of *S. bowleyana* is the roots with the compounds of phenols and terpenoids [6]. Clinically, it was used to treat irregular menstruation, amenorrhea dysmenorrhea, the pain of bone node, swelling pain of chest and flank, insomnia, angina, caruncle, neurasthenia, rheumatism, chronic hepatitis, ulcer of stomach, and duodenum [7]. *S. splendens* was initially found in Brazil and widely cultivated as horticultural plants in China. It acts as beautiful ornamental flowers and can be used to clear heat and cooling blood, eliminate swelling, and relieve pain [8]. Volatile oils are extracted from leaves of *S. officinalis*, which contain familiar constituents of Carnocera, pinene, cajuputole, borneol, and camphor. It has multiple functions such as anti-corrosion, antibacterial, anti-inflammatory, calming the nerves, and beautifying skin [9, 10].

Chloroplast is the essential organelle in the plants. The chloroplast genome contains a variety of genetic genes closely related to photosynthesis [11], evolution [12], and applications in genetic engineering [13]. In general, the chloroplast genome encodes more than 120 genes. These genes can be divided into three types [14] related to transcription and translation, photosynthesis, the biosynthesis of amino acids and fatty acids. The genes distributed in the LSC and SSC regions are mainly related to photosynthetic systems I(PSA) and systems II(PSB). They also include genes encoding Rubisco large subunit(*rbc*L) and small subunit gene(*rbc*S), tRNA gene(*t*RNA), ATP enzyme gene(ATP), NADH plastid masking oxidoreductase gene(NADH), and RNA polymerase gene(RPO) [15]. The genes distributed in the IRs region are mainly of the genes encoding rRNA(RPS), including 16S and 23S genes, the intermediate genes being separated by encoding 4.5S rRNA, 5S rRNA and 2tRNA genes, and some genes with unknown gene function [16].

The genes from chloroplast genomes can be used in species identification [17], phylogenetic evolution [18], genetic transformation [19], and molecular breeding of medicinal plants [20], providing basic data for resource identification and conservation. The sequences in the chloroplast genomes of medicinal plants, such as *psb*A-*trn*H, *mat*K, and *rbc*L, have been widely used for DNA molecular identification, and have now been developed to the analysis of polymorphic locus combinations of multiple genes and gene spacers [21]. Until now, the chloroplast genomes of the 14 *Salvia* species in the Lamiaceae family have been informed, including *S. miltiorrhiza* [22–24], *S. przewalskii* [24, 25], *S. bulleyana* [24], *S. japonica* [24], *S. plebeia* [26], *S. yunnanensis* [27], *S. miltiorrhiza f. alba* [28], *S. yangii* [29, 29], *S. chanryoenica* [30], *S. tiliifolia* [31], *S. hispanica* [32], *S. daiguii* [33], *S. leucantha* [34], and *S. trijuga* [35].

Compared with the diversification of nuclear and mitochondrial genome, the comprehensive development of chloroplast genomes could provide the basic database for further exploration regarding the characteristics, genetic evolution, and chemicals [36]. Therefore, we sequenced and analyzed the chloroplast genomes of three *Salvia* species for the first time so as to make an invaluable bioexploration between the evolutionary differences and similarities in the Lamiaceae family.

Results

Morphological characteristics of the three *Salvia* species

The three of *Salvia* species have the common specifications in the Lamiaceae family: quadrangular stem, opposite leaves, corolla flower lip, and 4 nutlets. However, they have the obvious distinction from the phenotype of flower colors varying from pink, purple(*S. bowleyana* and *S. officinalis*) to red(*S. splendens*). Moreover, the three *Salvia* species are perennial herbs with oblong or oval leaves, cymose inflorescences and nutlets. Nevertheless, for *S. bowleyana*, the leaves are glabrous on both sides, only the veins are slightly pilose, and the top of the fruit is hairy(Figure 1A). For *S. splendens*, while the stems, leaves on both sides, and petioles are not glabrous with glandular spots below. The fruits have the irregular folds at the top, and narrow wings at the edge(Figure 1B). For *Salvia officinalis*, the stems, many branches, leaf surfaces, and petioles are covered with white short villi. The fruits of it is smooth and hairless(Figure 1C) [1].

Gene compositions comparison of 23 *Salvia* species

Schematic representions of *S. bowleyana, S. splendens,* and *S. officinalis* chloroplast genomes are shown in Figures 2(A, B, and C), respectively. The total assembled length of them were 151387 bp, 150604 bp, and 151163 bp, respectively. The lengths of large-single copy(LSC), small-single copy(SSC), and dual inverted repeat regions in the three chloroplast genomes were 82772 bp, 17573 bp, and 51042 bp for *S. bowleyana*; 82181 bp, 17857 bp, and 50566 bp for *S. splendens*; 82429 bp, 17510 bp, and 51224 bp for *S. officinalis*. The GC contents of the three chloroplast genomes were 38.01%, 38.04%, and 38.04%, separately(Table 1, Additional file 1, Table S1).

The chloroplast genomes of *S. bowleyana*, *S. splendens*, and *S. officinalis* contained 131, 130, and 131 genes, respectively, including 80, 79, and 80 protein-coding genes, 36 tRNA genes, and 8 rRNA genes (Additional file 1, Table S1). There are 14 PCGs(*rps*12(×2), *rps*7(×2), *rpl*2(×2), *rpl*23(×2), *ndh*B(×2), *ycf*2(×2), and *ycf*15(×2)), 14 tRNA genes(*trn*A-UGC(×2), *trn*E-UUC(×2), *trn*M-CAU(×2), *trn*L-CAA(×2), *trn*N-GUU(×2), *trn*R-ACG(×2), *trn*V-GAC(×2)), and 8 rRNA genes(*rrn*16S(×2), *rrn*23S(×2), *rrn*4.5S(×2), *rrn*5S(×2)) located in the both IRa and IRb regions(Table 1), respectively. Among the three genomes, twenty-two genes commonly exhibited introns, of which seven tRNA genes(*trn*K-UUU, *trn*L-UAA, *trn*C-ACA, *trn*E-UUC(×2), *trn*A-UGC(×2)) and twelve cis-splicing CDS genes(*rps*16, *atp*F, *rpo*C1, *ycf*3, *clp*P, *pet*D, *rpl*16, *rpl*2(×2), *ndh*B(×2), *ndh*A) had a single one intron. In particular, the three genes had one intron in the special species, of which both genes *trn*T-CGU and *pet*B are identified in the species of *S. bowleyana* and *S. splendens.* Whereas protein-coding gene *pet*B was only showed in the *S. officinalis.* Especially, two CDS genes of *ycf*3 and *clp*P displayed two introns and three exons(Table 2, additional 7, Figure S1). Additionally, the containing intron gene *trn*K-UUU, making up the *mat*K, had the largest intron in the three chloroplast genomes of *Salvia* species(2522bp, 2494bp, and 2517bp, respectively). Except for the plants of Pteridophyta and parasitic species, the chloroplast of land plants commonly contain the *mat*K mature enzyme gene in the intron of the lysine tRNA-K(UUU) gene [37-39], which acts as a splicing factor for introns of the highly structured ribozyme group II [40,41]. Furthermore, the three segments of *rps*12 genes were located at the region of LSC , IRa, and IRb of the chloroplast genomes, respectively. The *rps*12 gene is splitted by two introns; one intron between exon 2 and 3 is 528 bp in length, another intron between exon 1 and 2 is about 28 kb in length(table 2, additional 8, Figure S2). The latter intron is trans-spliced to produce mature rps12 mRNA (additional 8, Figure S2) [42]. The exon1 and the two copies of exons are trans-spliced together to form two transcripts. The arrow indicated the sense direction of the genes (additional 7,8, Figure S1, S2).

Among the 23 *Salvia* species, the lengths of the total genome, LSC, SSC, IR, protein-coding genes, tRNA genes, rRNA genes, and Non-coding region varied from 150604 bp to 153995 bp, from 82129 bp to 84775 bp, from 17464 bp to 17875 bp, from 25283 bp to 25815 bp, from 78558 bp to 80754 bp, from 2724 bp to 2815 bp, from 9046 bp to 9396 bp, and from 58691 bp to 62823 bp. The percentage of GC contents for the total genome, LSC, SSC, and IRs regions diversified from 37.94% to 38.05%, from 36.07% to 36.23%, from 31.63% to 32.07%, and from 43.06% to 43.20%. The gene numbers of the total genes, protein encoding genes, and tRNA genes ranged from 130 to 133, from 85 to 88, and from 36 to 37, respectively. The chloroplast genomes in all 23 *Salvia* species encoded two copies of *rrn*16S, *rrn*23S, *rrn*4.5S, *rrn*5S (Additional file 1, Table S1).

Table 1. Functional genes comparison of *Salvia bowleyana, Salvia splendens, Salvia officinalis* chloroplast genomes

Species/items		*S. bowleyana*	*S. splendens*	*S. officinalis*
Gene function	**Gene type**	**Gene name**		
tRNA	tRNA genes	36 *trn* genes (include one intron in 8 genes)	36 *trn* genes (include one intron in 8 genes)	36 *trn* genes (include one intron in 8 genes)
Photosynthesis	Subunits of ATP synthase	*atp*A, *atp*B, *atp*E, *atp*F, *atp*H, *atp*I		
	Subunits of photosystem Ⅰ	*psa*A, *psa*B, *psa*C, *psa*I, *psa*J		
	Subunits of photosystem Ⅱ	*psb*A, *psb*B, *psb*C, *psb*D, *psb*E, *psb*F, *psb*I, *psb*J, *psb*K, *psb*L, *psb*M, *psb*N, *psb*T, *psb*Z, *ycf*3		
Gene expression	Ribosomal RNAs	*rrn*16*s*[a], *rrn*16*s*[b], *rrn*23*s*[a], *rrn*23*s*[b], *rrn*4.5*s*[a], *rrn*4.5*s*[b], *rrn*5*s*[a], *rrn*5*s*[b]		
	DNA dependent RNA polymerase	*rpo*A, *rpoB*, *rpo*C1, *rpo*C2		
	Small subunit of ribosome	*rps*11, *rps*12[L], *rps*12[a], *rps*12[b], *rps*14, *rps*15, *rps*16, *rps*18, *rps*19, *rps2*, *rps*3, *rps*4, *rps*7[a], *rps*7[b], *rps*8		
	Large subunit of ribosome	*rpl*14, *rpl*16, *rpl*2[a], *rpl*2[b], *rpl*20, *rpl*22, *rpl*23[a], *rpl*23[b], *rpl*32, *rpl*33, *rpl*36		
	Subunits of NADH-dehydrogenase	*ndh*A, *ndh*B[a], *ndh*B[b],,*ndh*C, *ndh*D, *ndh*E, *ndh*F, *ndh*G, *ndh*H, *ndh*I, *ndh*J, *ndh*K		
	Subunits of cytochrome b/f complex	*pet*A, *pet*B, *pet*D, *pet*G, *pet*L, *pet*N		
	Ribulose diphosphate carboxylase subunit	*rbc*L		
Other genes	Subunit of Acetyl-CoA-carboxylase	*acc*D		
	C-type cytochrome synthase	*ccs*A		
	Protease	*clp*P		
	Translation initiation factor	*inf*A		
	Mature enzyme	*mat*K		
	Envelope membrane protein	*cem*A		
Unknown functions	Conservative open reading frame	*ycf*1[s-b], *ycf*2[a], *ycf*2[b], *ycf*15[a], *ycf*15[b], *ycf*4		

Note: L: LSC region; a: IRa region; b: IRb region; s-b: Across the SSC and IRb regions.

Table 2. The lengths of introns and exons for the splitting genes in the *S. bowleyana*, *S. splendens*, *S. officinalis* chloroplast genomes

Gene name	strand	initial position- final position			Length(bp)												
		S. bowleyan	*S. splendens*	*S. officinalis*	The first exon			The first intron			The second exon			The second intron			Th
		A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A
trnK-UUU	-	1672-4266	1684-4250	1703-4292	37	37	37	2522	2494	2517	36	36	36				
rps16	-	4835-5945	4819-5917	4863-5972	40	40	40	874	862	873	197	197	197				
trnT-CGU	+	9001-9755	8765-9528	/	35	35	/	677	686	/	43	43	/				
trnS-CGA	+	/	/	8621-9377	/	/	32	/	/	665		/	60				
atpF	-	11742-12989	11506-12764	11353-12606	145	145	145	693	704	699	410	410	410				
rpoC1	-	20712-23525	20528-23339	20399-23215	430	430	430	759	757	762	1625	1625	1625				
ycf3	-	41963-43894	41526-43464	41641-43591	129	129	129	696	702	706	228	228	228	726	727	735	15
trnL-UAA	+	46799-47338	46350-46917	46202-46773	35	35	35	455	483	487	50	50	50				
trnC-ACA	-	50870-51518	50236-50881	50440-51087	38	38	38	555	552	554	56	56	56				
rps12^{L}		68691-68804	68105-68218	68355-68468	114	114	114										
clpP	-	68928-70839	68342-70250	68591-70509	71	71	71	692	703	711	294	294	294	629	615	617	22
petB	+	73746-75096	73171-74533	/	6	6	/	703	715	/	642	642	/				
petD	-	75290-76492	74721-75904	74979-76169	8	8	8	720	701	708	475	475	475				
rpl16	-	79937-81217	79325-80600	79599-80867	9	9	9	873	868	861	399	399	399				
rpl2	-	82875-84357	82266-83757	82532-84019	391	391	391	658	667	663	434	434	434				
ndhB	+	93058-95211	92464-94617	92711-94918	721	721	775	675	675	675	758	758	758				
rps12^{b}		96061-96844	96018-96260	95714-96507	114	114	114	/	/	/	232	243	232	528	/	538	26
trnE-UUC	+	100535-101546	99979-100997	100210-101229	32	32	32	940	947	948	40	40	40				
trnA-UGC	+	101611-102478	101062-101938	101294-102171	37	37	37	795	804	805	36	36	36				
ndhA	-	117349-119425	116488-118588	117038-119137	553	553	553	985	1009	1008	539	539	539				
trnA-UGC	-	131682-132549	130848-131724	131422-132299	37	37	37	795	804	805	36	36	36				
trnE-UUC	-	132614-133625	131789-132807	132364-133383	32	32	32	940	947	948	40	40	40				
rps12^{a}		137316-138099	136526-136768	137086-137879	114	114	114	/	/	/	232	241	232	528	/	528	26
ndhB	+	138949-141102	138169-140322	138675-140882	721	721	775	675	675	675	758	758	758				
rpl2	+	149803-151285	149029-150520	149574-151061	391	391	391	658	667	663	434	434	434				

Note: "+" indicates a positive chain; "-" indicates a negative chain; A: *S. bowleyan*; B: *S. splendens*; C: *S. officinalis*. L: LSC region; a: IRa region; b: IRb region.

Gene loss analysis of the chloroplast genomes from 41 species in the Lamiaceae family

The gene losses of chloroplast genomes were analyzed in the 41 species of Lamiaceae family originated from the phylogenetic tree(Table 3 and Figure 6). These species originated from 8 genera(*Salvia*, *Rosmarinus, Agastache, Dracocephalum, Ajuga, Leonurus, Elsholtzia, Caryopteris)* of the Lamiaceae family. In the dual IR regions of chloroplast genomes, the one of *rpl*20 gene was stable and found in all 41 species; however, the another one only was found in *D. heterophyllum*. Therefore, the intact *rpl*20 gene often can be used to the molecular signature gene in the angiosperm [43]. the In addition, One of the *ycf*1 gene is across the SSC and IRb regions, the another pseudogene one is acorss the SSC and IRa regions. The losses of the first *ycf*1 gene was observed in five chloroplast genomes of *A. campylanthoides*, *A. ciliata*, *A. decumbens*, *A. lupulina*, *A. nipponensis*. The losses of the second one was not found in the chloroplast genomes of twenty-eight species except the twelve chloroplast genomes from the six *Salvia* genus(*S. digitaloides, S. daiguii, S. meiliensis, S. chanryoenica, S. yangii,* and *S. nilotica*)*, A. rugosa,* the four *Dracocephalum* genus*(D. heterophyllum, D. taliense, D. tanguticum,* and *D. moldavica*), and *L. japonicas.* As reported, in total of 420 species, 357 species could be distinguished using *ycf*1 by means of specific primers designed for the amplification of these regions [44]. Moreover, the losses of *ycf*15 genes was occurred in five chloroplast genomes(*S. hispanica*, *S. tiliifolia, S. chanryoenica, A. forrestii*, and *E. densa*). Although the gene function of *ycf*15 genes is unknown, the transcriptome analyses of *Camellia genus* revealed that ycf15 was transcribed as precursor polycistronic transcript which contained *ycf*2, *ycf*15 and antisense *trn*L-CAA [45].Furthermore, the six genes in LSC region, e.g. *pet*N, *acc*D, *rps*2, *rps*16, *rps*18, and *rps*19 were absent in the chloroplast genomes of *C. trichosphaera, R. officinalis*, *D. moldavica, E. densa, D. heterophyllum*, and *L. japonicus*, *respectively*. Whereas in the SSC region, the losses of the *rpl*32 and *ndh*D genes were only found in *S. splendens,* and *C. mongholica* chloroplast genomes, respectively. Surprisingly, the loss of *rpl*32 gene can be transferred to the nucleus from the *Euphorbia schimperi* chloroplast genome and this can be verified through the method of being sequenced in the nuclear transcriptome of *E. schimperi* [46]. The type of gene losses was mostly affirmed to be consistent with the topology of the evolutionary tree.

Table 3 Gene losses of the different regions in the 41 chloroplast genomes from the Lamiaceae family.

Genus	Name of species	the genes of IR region				The genes of LSC region						The genes of SSC region	
		*rpl*20_copy	*ycf*1	*ycf*1_copy	*ycf*15	*pet*N	*acc*D	*rps*2	*rps*16	*rps*18	*rps*19*	*rpl*32	*ndh*D
Salvia	*S. bowleyana*	-	+	-	+	+	+	+	+	+	+	+	+
	S. splendens	-	+	-	+	+	+	+	+	+	+	-	+
	S. officinalis	-	+	-	+	+	+	+	+	+	+	+	+
	S. bulleyana	-	+	-	+	+	+	+	+	+	+	+	+
	S. digitaloides	-	+	+	+	+	+	+	+	+	+	+	+
	S. japonica	-	+	-	+	+	+	+	+	+	+	+	+
	S. plebeia	-	+	-	+	+	+	+	+	+	+	+	+
	S. przewalskii	-	+	-	+	+	+	+	+	+	+	+	+
	S. yunnanensis	-	+	-	+	+	+	+	+	+	+	+	+
	S. miltiorrhiza	-	+	-	+	+	+	+	+	+	+	+	+
	S. daiguii	-	+	+	+	+	+	+	+	+	+	+	+
	S. miltiorrhiza f.*alba*	-	+	-	+	+	+	+	+	+	+	+	+
	S. meiliensis	-	+		+	+	+	+	+	+	+	+	+
	S. hispanica	-	+	-	-	+	+	+	+	+	+	+	+
	S. merjamie	-	+	+	+	+	+	+	+	+	+	+	+
	S. sclarea	-	+	-	+	+	+	+	+	+	+	+	+
	S. petrophila	-	+	-	+	+	+	+	+	+	+	+	+
	S. tiliifolia	-	+	-	-	+	+	+	+	+	+	+	+
	S. chanryoenica	-	+	+	-	+	+	+	+	+	+	+	+
	S. yangii	-	+	+	+	+	+	+	+	+	+	+	+
	S. Prattii Hemsl.	-	+	-	+	+	+	+	+	+	+	+	+
	S. roborowskii	-	+	-	+	+	+	+	+	+	+	+	+
	S. nilotica	-	+	+	+	+	+	+	+	+	+	+	+
Rosmarinus	*R. officinalis*	-	+	-	+	+	-	+	+	+	+	+	+
Agastache	*A. rugosa*	-	+	+	+	+	+	+	+	+	+	+	+
Dracocephalum	*D. heterophyllum*	+	+	+	+	+	+	+	+	-	+	+	+
	D. taliense	-	+	+	+	+	+	+	+	+	+	+	+
	D. tanguticum	-	+	+	+	+	+	+	+	+	+	+	+
	D. moldavica	-	+	+	+	+	+	-	+	+	+	+	+
Ajuga	*A. forrestii*	-	+	-	-	+	+	+	+	+	+	+	+
	A. campylanthoides	-	-	-	+	+	+	+	+	+	+	+	+
	A. ciliata	-	-	-	+	+	+	+	+	+	+	+	+
	A. decumbens	-	-	-	+	+	+	+	+	+	+	+	+
	A. lupulina	-	-	-	+	+	+	+	+	+	+	+	+
	A. nipponensis	-	-	-	+	+	+	+	+	+	+	+	+
Leonurus	*L. japonicus*	-	+	+	+	+	+	+	+	+	-	+	+
Elsholtzia	*E. densa*	-	+	-	-	+	+	+	-	+	+	+	+
Caryopteris	*C. trichosphaera*	-	+	-	+	-	+	+	+	+	+	+	+
	C. mongholica	-	+	-	+	+	+	+	+	+	+	+	-
	C. incana	-	+	-	+	+	+	+	+	+	+	+	+

C. forrestii	-	+	-	+	+	+	+	+	+	+	+	+	

Note: The +/- refers to the presence/absence of a gene in each species that do not have the gene. "+": presence; " − " absence; **rps*19 is across the area of LSC and IRb.(add family and order information)

Analysis of SSR Polymorphism in the 23 *Salvia* chloroplast genomes

Repeat sequences have been commonly used as the genetic markers to understand and evolution of the genus in the same family [16]. Scattered(interspersed) repetition and tandem repetition sequences consisting of simple tandem repeats(SSR) were analyzed in the 23 *Salvia* chloroplast genomes (Additional file 2, Table S2, Fig. 3). We analyzed the content and percentage of SSR sequences in the 23 *Salvia* species. The results showed that 16, 12, 10 SSR contained "A" as the repeat unit and 18, 14, 14 SSR contained "T" as the repeat unit among the total 34, 26, 24 mononucleotide repeats (Additional file 2, Table S2A, S2B) in the chloroplast genomes of *S. bowleyana*, *S. splendens*, and *S. officinalis*, respectively. Moreover, the mononucleotide numbers of "A" and "T" as the repeat unit have an obvious difference. From the statistical results, the number Poly A and Poly T repeats varied from 6(*S. yangii*) to 16(*S. bowleyana* and *S. miltiorrhiza f. alba*), from 9(*S. plebeia*) to 21(*S. prattii*). While the rare numbers of Poly C and Poly G repeats only were found in the chloroplast genomes of *S. hispanica*, *S. plebeia*, and *S. meiliensis* [47]. One SSR with "AT" as the repeat unit was found in the eight *Salvia* chloroplast genomes of *S. splendens, S. digitaloides, S. daiguii, S. hispanica, S. tiliifolia, S. chanryoenica, S. prattii,* and *S. roborowskii*. Di-nucleotide SSR contained "TA" as the repeat unit in twelve chloroplast genomes of *S. bowleyana, S. bulleyana, S. przewalskii, S. yunnanensis, S. miltiorrhiza f. alba, S. chanryoenica, S. prattii, S. roborowskii, S. splendens, S. daiguii, S. hispanica,* and *S. tiliifolia,* respectively. Nevertheless, one trinucleotide SSR with "AAT" as the repeat unit was found in the chloroplast genome of *S. yunnanensis (*Additional file 1, Table S2). The mononucleotide of repeat unit is the most abundant type of the SSR repeats and it accounted for the proportion from 88% to 100% through comprehensive statistics of chloroplast genomes in the 23 *Salvia* species.

Repeat sequences analysis in the chloroplast genomes of 23 *Salvia* species

Except for the SSR analysis of 23 *Salvia* chloroplast genome, 29 tandem repeats by each species were identified for all the four kinds of tandem repeats, including the forward repeats, reverse repeats, palindromic repeats, complement repeats in the chloroplast genomes of *S. bowleyana (*11 forward repeats, 3 reverse repeats, and 15 palindromic repeats), *S. splendens (*11 forward repeats, 4 reverse repeats, and 14 palindromic repeats) and *S. officinalis (*10 forward repeats, 5 reverse repeats, and 14 palindromic repeats), respectively. The most number of repeats type were forward repeats and palindromic repeats. While the number of reverse repeats and complement repeats is less and the latter is only found in the six chloroplast genomes, including *S.przewalskii*, *S.daiguii*, *S.meiliensis*, *S.merjamie*, *S.yangii*, and *S.nilotica*. The comparison of the number of the predicted tandem repeats are shown in Table S3, Table S4 (Additional file 3,4) and Figure 2C.

Among the 23 *Salvia* chloroplast genomes of the interspersed repeats, the number of palindromic and direct repeats varied from 14 (*S. merjamie, S. sclarea,* and *S. daiguii*) to 26 (*S. miltiorrhiza*, *S. petrophila, S. prattii, S. roborowskii,* and *S. splendens*). The number of tandem repeats will be reduced by more than half and diversified from 6 (*S. bowleyana, S. splendens, S. plebeia, S. miltiorrhiza,* and *S. miltiorrhiza* f. *alba*) to 24 (*S. japonica*) while the similarity among the repeat unit sequences > = 90%. The e-values of interspersed repeats varied from 7.65E-23 to 6.07E-04 (Additional file 5, Table S5). In this study, forty-seven interspersed repeat (25 palindromic repeats and 22 direct repeats), forty-nine interspersed repeats (23 palindromic repeats and 26 direct repeats), forty interspersed repeats (20 palindromic repeats and 20 direct repeats) were identified in the chloroplast genomes of *S. bowleyana, S. splendens, S. officinalis*, respectively, with the length of repeat units 1, 2 being between 30 bp and 63 bp (Additional file 5, Table S5).

Structures of the IR boundaries and gene features from 23 *Salvia* species

The IR boundaries' structure was analyzed in the 23 *Salvia* chloroplast genomes of Lamiaceae family. From the analysis, six distinct genes *rpl*22, *rps*19, *rpl*2(×2), *ycf*1, *ndh*F, and *psb*A were the most explicitly found in the diverse regions or at the border regions of 23 chloroplast genomes(Figure 4). Furthermore, variation range of these gene lengths is similar and do not exceed 2%. The genes of *rpl*22 and *psb*A were located in the LSC region, whereas *rpl*2 genes were located in the two IR regions in these species. The one of *rps*19 genes was located at the border area of LSC and IRb in all species. In addition, a small fragments of the *rps*19 genes(*rps*19 pseudogene) were found at the border regions of the LSC and IRa in fourteen chloroplast genomes of *S. bulleyana, S. digitaloides, S. japonica, S. plebeia, S. przewalskii, S. miltiorrhiza, S. daiguii, S. miltiorrhiza f.alba, S. meiliensis, S. petrophila, S. yangii, S. nilotica, S. prattii, S. roborowskii*, in consistent with the existing studies [28]. In contrast, the *ycf*1 genes were traversed the border regions of SSC and IRb in all 23 species, while *ycf*1 gene fragments(*ycf*1 pseudogene) were found at the border regions of SSC and IRa in six *Salvia* chloroplast genomes(*S. merjamie, S. digitaloides, S. daiguii, S. chanryoenica, S. nilotica,* and *S.yangii*) . Besides, *ndh*F genes were located at the border regions of IRa and SSC in all 23 species. The IRa/LSC boundary positions were located on the *trn*H genes in the five chloroplast genomes of *S. chanryoenica*, *S.splendens, S. nilotica, S. yangii, S. tiliifolia.* Especially, a little bit fragment of *trn*N gene located in the IRb region of the *Salvia splendens* chloroplast genome(Fig. 4), popularly found in the *Cymbidium* genus among the photosynthetic Orchids [48].

The discrepancy of the 23 *Salvia* chloroplast genomes

The structure of chloroplast genome is stable and the DNA is informative. The medicinal plants can be accurately identified and distinguished by the comparison of barcodes from the whole chloroplast genome [49]. The sequences of chloroplast genomes in the 23 *Salvia* species were analyzed using mVISTA, and the alignments were visualized with the *Salvia bowleyana* chloroplast genome as the reference genome (additional 9, Figure S3). We found the sequences of 23 *Salvia* chloroplast genome were mostly identical conserved except of the three variable areas located in the intergenic regions of LSC region. The first one is the IGS *rps*16-*trn*Q-UUG found in the nine *Salvia* chloroplast genomes(*S. officinalis, S. japonica, S. sclarea, S. meiliensis, S. hispanica, S. tiliifolia, S. yangii, S. splendens, S. nilotica*) (panel A). The second one is the IGS *trn*L-UAA-*trn*F-GAA varied in the chloroplast genome of *S. chanryoenica*(panel B). The last one is the IGS *trn*M(cau)-*atp*E diversified in the three chloroplast genome of *S. chanryoenica, S. hispanica* and *S. japonica*(panel C).

Identification of hypervariable regions

It is significant to develop molecular markers in the chloroplast genomes of plants by identifying the highly variable sites [50]. We analyzed the genetic distance among the intergenic spacer regions(IGS) in the chloroplast genomes of 23 *Salvia* species. The results showed that K2p distances of 91 IGS regions ranged from 0.00 to 20.41(Additional file 6, Table S6). Among them, 30 IGS regions had K2p distances varying from 3.65 to 20.41(Figure 5). Particularly, five IGS regions had the higher K2p values diversified from 6.36 to 20.41, which they were *the* regions of *trn*L-UAG-*ccs*A(20.41), *rps*16-*trn*Q-UUG(13.42), *ccs*A-*ndh*D(7.98), *rps*15-*ycf*1(6.63), and *ndh*E-*ndh*G(6.36). In general, the large K2p distances indicate a high degree of sequence divergences. Thus, these five regions of IGS can be suitable candidates for developing molecular markers in the 23 *Salvia* species, whereas, two markers derived from the *pet*N-*psb*M and *psa*J-*rpl*33 IGS regions that successfully distinguished the five *Alpinia* species [51, 52].

Identification and comparison of the Genus-specific DNA barcodes

Primers can be designed from highly variable intergenic spacer sequences for PCR amplification. Then, we can distinguish the 23 *Salvia* species in the Lamiaceae family by sequence alignment and analysis using ecoPrimers software. After comparison, the two conservative intervals can be amplified through the designed PCR amplification primers to distinguish the 23 *salvia* genus. The primers and amplified sequences are shown in Table 4. Surprisingly, the two pairs of primers can be used to amplified the sequences of *trn*M-CAU-*atp*E and *ccs*A-*ndh*D after comparison between the *Salvia* chloroplast genomes and the BlastN database.

Furthermore, the alignment results based on the blast database indicate that the two pair primers can also be especially suitable to distinct other species [53], e.g. *Scutellaria* genus(Lamiaceae), *Camellia* genus(Theaceae), *Styrax* genus(Styracaceae), *Melissa* genus(Lamiaceae), *Eucalyptus* genus(Myrtaceae), etc.

Table 4 The conserved sequences for designing primers for the amplification of the DNA barcodes to distinguish 23 *Salvia* species in the Lamiaceae family

No	species	conserved sequences for designing Forward primers	conserved sequences for designing Reverse primers
1	23 *Salvia* species	TTTTCCCCTTCCTACCCC	AAAAAAAGATGTTGCGGAGACAGGATTTGAACCCGTGACCTCAAGGTTATGAGCC
2		TTACATAGTTATGGTTCATTTACATTAACATCTAATTAAAT	TTTTTTCATTGTACAACGAAC

Phylogenetic analysis

The sequences of chloroplast genomes are the valuable database for the research of the evolutionary relationship in the plants [17]. To determine the phylogenetic positions of the three *Salvia* species in the Lamiaceae family, 80 proteins sequences were extracted using the PhyloSuite software from the 43 chloroplast genomes in the species (Additional file 1, table S1). Among them, 25 shared CDS proteins sequences were found present in 43 species, including *rpl*14, *rpl*33, *rpl*36, *rps*7, *rps*14, *psb*B, *psb*C, *psb*D, *psb*E, *psb*F, *psb*N, *psa*B, *psa*C, *psa*I, *pet*A, *pet*G, *pet*L, *ndh*C, *ndh*G, *cem*A, *atp*A, *atp*B, *atp*H, *atp*I, and *ycf*4 genes. Using *L. chuanxiong(*Apiaceae family) and *P. notoginseng* (Araliaceae family) as the outgroups, a Maximum likelihood (ML) phylogenetic tree was generated based on the above-described data of whole chloroplast genome. The phylogenetic tree showed that 41 species including 37 species of Lamiaceae family and four species of Verbenaceae family were clustered together. The total branches of 41 species were divided into 6 obvious clades, that is to say, five species including *Dracocephalum* species(*D. heterophyllum*, *D. Taliense, D. tanguticum,* and *D. moldavica*) and *A. rugose* were clustered into one branch; in contrast, 23 *Salvia* species and one *Rosmarinus* species(R. *officinalis*) were clustered into one branch with six subbranches(Figure 6). In addtion, six species from *Ajuga* genus and four species from *Caryopteris* genus were clustered into the other two branches, respectively. Single species of *L. japonicus* and *Elsholtzia densa* were gathered into one branch, partly. Whereas, the species of outgroups were more distantly related from other species. The ML bootstrap shows strong support with the bootstrap values of 100% for eight nodes. The phylogenetic results resolved 26 nodes with bootstrap support values of 54-100 and that of 17 nodes were ≥ 74% (Figure 6).

Discussion

In the three *salvia* chloroplast genomes of *S. bowleyana, S. splendens*, and *S. officinalis*, the total number of protein-coding genes and CDS genes differ justly one short(*S. splendens*), tRNA and rRNA genes are the same as the most of other *Salvia* species, which illustrates that the chloroplast genomes are conserved in *Salvia*. The selected 41 species from the Lamiaceae family and the two outgroup species(*L. chuanxiong* and *P. notoginseng*) possessed similar pharmcological effects, such as, promoting blood circulation for removing blood stasis, increasing coronary flow, improving microcirculation, protecting the heart, improving the body hypoxia resistance, anti-hepatitis, antitumor and antiviral [54]. Chloroplasts play an irreplaceable role in the formation of chemicals and the development of phenotype due to the genes from nuclear, mitochondrial genomes. However, the variability of the nuclear genome was found to be higher than that of the chloroplast genome and mitochondrial genome, as the reported from average genetic distance among all the strains of CWR and cultivated rice [55]. Therefore, it is much indispensable to analyze the chemical composition and genetic divergence combined with a variety of genomics.

The *mat*K gene in the chloroplast genome has been commonly used in the plant identification and systematics to construct the phylogeny above the family level as a result of the rates, patterns, and types of nucleotide substitutions [56]. As reported, the complete *mat*K sequences from 11 seed plants and liverworts, and nine partial sequences representing the families of monocot(such as Poaceae, Joinvilleaceae, Cyperaceae, and Smilacaceae) were analyzed. Results showed that conserved 3' region and the less conserved 5' region of *mat*K gene can be used at different taxonomic level [57]. Additionally, three regions of functional importance of *mat*K gene have been identified and they showed highly conserved secondary structure, of which supported the putative

function of *mat*K as a group II intron maturase [58]. Furthermore, *mat*K gene has been proposed to help with the removal of seven distinct chloroplast group IIA introns through the increasing *mat*K messenger RNA in mature tissue, which located in the precursor of ribonucleic acid, acting as an essential element for chloroplast function [41, 59]. Moreover, *mat*K protein increases efficiency of group IIA intron self-splicing for the second intron of *rps*12 [39]. The *rps*12 genes with intron-containing and intron-less were significant differences in the patterns and rates of nucleotide substitutions through the investigation of 16 complete fern plastome sequences [60]. Likely, one of the *ycf*1 genes also has the most variable region [44, 61]. On the contrary, *ycf*15 gene was found in the most of chloroplast genomes from the *Salvia* species, while can be identified that of *Wisteria floribunda* and *W. sinensis* in the Papilionoideae subfamily [62]. The structural changes of above genes together with the comparison of the chloroplasts regions can serve as a core barcode and have the important effects on evolutionary rates of land plants.

Similarly, it makes sense that the DNA sequences of the hypervariable regions and comparison of chloroplast genomes in three IGS regions of *rps*16-*trn*Q-UUG, *trn*L-UAA-*trn*F-GAA, and *trn*M(cau)-*atp*E can definitely be used to distinguish the ten *Salvia* species(*S. officinalis*, *S. japonica*, *S. sclarea*, *S. meiliensis*, *S. hispanica*, *S. tiliifolia*, *S. yangii*, *S. splendens*, *S. nilotica*, and *S. chanryoenica)*. The first IGS region has been found in the species of *Zingiber officinale* and *Cofeeae alliance* [63, 64]. The second one commonly occurred in the Angiosperm [65]. The last one diversified and some part of the oldest mtDNAs of *trn*V(uac)-*trn*M(cau)-*atp*E-*atp*B-*rbc*L transferred from cpDNA to mtDNA since the common ancestor in extant gymnosperms and angiosperms [66]. Therefore, the DNA barcodes for the identification and phytotaxonomy of genus *Salvia* species were potentially developed through the divergence region of IGS.

As reported, ninety-one taxa of EA *Salvia* were sampled and 34 taxa of *Salvia* were analyzed based on the DNA markers of internal transcribed spacer(ITS), external transcribed spacer(ETS), and four chloroplast sequence(*psb*A-*trn*H, *ycf*1-*rps*15, *trn*L-trnF and *rbc*L). All *Salvia* species native to East Asia formed a clade, and eight major subclades(A-G) were recognized [67].

The sequences of inverted repeat(IR) can complement a certain segment of the upstream sequence in downstream of the same DNA strand. It can then form a hairpin structure with a double helix stem and a single-stranded ring with a DNA double helix. The sequence between two reverse repeat units forms a single chain loop. Two copies are separated by a sequence or no interval sequence, which is in reverse series, and will form a specific palindrome sequence(P) [68]. Compared to the IRLC between the Papilionoideae subfamily and Lamiaceae family, they have the four common genes of *ndh*B, *rpl*23, *ycf*1, and *ycf*15.

In the IR regions, the genes of *ndh*B, *rpl*2, *rpl*23, *rps*7, *rps*12, *ycf*2 were present in the chloroplast genomes of 41 species, and these genes have a special function in the area of gene expressions. There are five hypothetical coding regions genes of *ycf*1, *ycf*2, *ycf*4, *ycf*15 and two open reading frames(ORF42 and ORF56), which are also found in the chloroplast genomes of the other species, such as *Clerodendranthus spicatus* [69]. Both genes *ycf*3 and *ycf*4 were present in the LSC region of the 41 species chloroplast genomes. The sequence of *ycf*3 is conserved in plants and contains three tetratrico-peptide repeats(TPR), which can act as the essential functions for the accumulation of the photosystem I(PSI) complex through a post-translational level [70, 71]. The *ycf*4 gene form modules that mediate PSI assembly and facilitates the integration of peripheral PSI subunits and LHCIs into the PSI reaction center subcomplex [72, 73].

Conclusions

We sequenced and acquired the complete chloroplast genomes of *S. bowleyana*, *S. splendens*, and *S. officinalis* using Illumina sequencing technology compared to the 23 *salvia* chloroplast genomes. These three species can be easily discriminated from the pheotype. Phylogenetic analysis showed that 23 *Salvia* species and one *Rosmarinus* genus(R. *officinalis*) were clustered into one branch with six subbranches, of which the three of *S. bowleyana, S. splendens*, and *S. officinalis* were included in the diverse branches. The sequences divergence found seven sites: IGS(*rps*16-*trn*Q-UUG), IGS(*trn*L-UAA-*trn*F-GAA), IGS(*trn*M-CAU-*atp*E), IGS(*trn*L-UAG-*ccs*A), IGS(*ccs*A-*ndh*D), IGS(*rps*15-*ycf*1), and IGS(*ndh*E-*ndh*G). And the sequences divergence had the higher variability and can be developed the DNA marker for the identification and phytotaxonomy of genus *Salvia* species. Overall, the data obtained will contribute to further development of the authentication, diversity, ecology, taxonomy, phylogenetic evolution and conservation of *Salvia* genus in China.

Materials And Methods

Plant photos and materials

Salvia bowleyana, S. splendens, and *S. officinalis* are the three characteristic plants from the *Salvia* genus of the Lamiaceae family(Figure 1). The photos of plants were provided and identified by Professor Peng LQ from the chuzhou Hospital of integrated traditional Chinese and Western medicine, Anhui Province. The three *S. bowleyana*, *S. splendens*, and *S. officinalis* plants were from Jiangsu nanjing Botanical Garden Mem, In the Civic Park of Guangdong, and Anhui chuzhou Mount langya, respectively. We collected the young leaves of *S. bowleyana, S. splendens* and *S. officinalis* from the Guangxi Medical Botanical Garden, Nanning, Guangxi, China(Geospatial coordinates: N22.859968, E108.383475) and dried by Silica gel immediately for total genomic DNA isolation. The voucher specimens were deposited at the Institute of Medicinal Plant Development under the voucher number: implad201910237, 201808155, 20170492, respectively(Contact person: HM Chen; Email: hmchen@implad.ac.cn).

DNA extraction and determination of DNA quality

Total genomic DNA was extracted from the dried leaves using the plant genomic DNA kit(Tiangen Biotech, Beijing, China). The DNA purity was detected by 1.0% agarose gel. Moreover, we detected the DNA concentration using the Nanodrop spectrophotometer 2000 [74].

Chloroplast genome sequencing, assembly, annotation, and manual curation

DNA extracts were fragmented for 300 bp short-insert library construction. The library was sequenced in pair-end model with the read length of 150 bp on an Illumina Hiseq 2500 platform [75]. The raw reads were filtered using Trimmomatic 0.35 with default parameters to remove adapters and low-quality bases [76]. The three chloroplast genomes were assembled using the NOVOPlasty(v 4.2) software [77] with the default parameters and the *rbc*L sequences as the

seed. After that, we annotated the genome using the CpGAVAS2 web service(http://www.herbalgenomics.org/cpgavas2/) [78]. The annotation errors were manually corrected using the Apollo software [79]. At last, the assembly and the annotation results of *S. bowleyana*, *S. splendens,* and *S. officinalis* were submitted to GenBank with the accession numbers: OM617845, OM617847, and OM617846, respectively.

Visualization and analysis of genome content, cis- and trans-splicing genes

We visualized the chloroplast genome structure, cis-splicing genes, and trans-splicing PCGs using CPGview-RSG software(http://www.herbalgenomics.org/cpgview/). The gene contents of 41 studied species (Additional file 1, Table S1) were analyzed including length of the complete geome sequences and the four regions, all genes, CDS, tRNAs, and rRNAs.

Repeat analysis

We annotated the repeat sequences using the CPGAVAS2 online tool(http://www.herbalgenomics.org/cpgavas2/) in the chloroplast genomes of *S. bowleyana, S. splendens,* and *S. officinalis*. The simple sequence repeats(SSR) of 23 *Salvia* species were identified using MISA software(http://pgrc.ipk-gatersleben.de/misa/) [80], also called the microsatellite sequence. The search parameter was set as: the minimum number of repeats with the bases number being mononucleotide, dinucleotide, trinucleotide, tetranucleotide, pentanucleotide, hexanucleotide, and hexagenucleotide are 10,5,4,3,3,3 and 3, respectively. The minimum distance between the 2 SSR was set to 100 bp. If the distance is less than 100 bp, the two SSR are treated as a composite microsatellite. TRS of the 23 *Salvia* chloroplast genome were predicted using the TRF software [81]. IRS were predicted using the REPuter program(https://bibiserv.cebitec.uni-bielefeld.de/reputer), with the parameters as follows: Maximum Computed Repeats = 30 and Minimal Repeat Size = 8) [82]. The comparison of chloroplast genome was conducted using VMATCH software(Professor Stefan Kurtz, Computer Science at the Center for Bioinformatics, University of Hamburg, Germany).

Comparative Genomic Analysis

We downloaded 40 chloroplast genomes sequences from the GenBank database including 38 species from the Lamiaceae family and two outgroups(*Ligusticum chuanxiong* from the Apiaceae family and *Panax notoginseng* from the Araliaceae family, for further analysis. The boundaries of the LSC, SSC, and IR regions boundary of chloroplast genomes from 23 *Salvia* species were visualized using the IR scope software(https://irscope.shinyapps.io/irapp/) [83]. Henceforth, we analyzed the characteristic genes including the diverse areas. The chloroplast genome sequences of 23 species from *Salvia* genera were compared with the annotated *S. bowleyana* chloroplast as the reference using the mVISTA program in a Shuffle-LAGAN mode with default parameters(Rank VISTA probability threshold=0.5) [84-85]. The genetic distances of intergenic spaces(IGS) from the chloroplast genomes of 23 *Salvia* species were calculated by using the distmat program from EMBOSS(v6.3.1) [86] with the Kimura 2-parameters(K2p) evolutionary model [87].

The identification of Genus-specific DNA barcode sequences

To discover the DNA barcode sequences that can distinguish the 23 *Salvia* species, we analyzed the PCR amplification primers from their chloroplast genome sequences using ecoPrimers software [88]. Moreover, the sequences of two pairs primers have been compared to the other species through the CBI Multiple Sequence Alignment Viewer(Version 1.21.0, Max Seq Difference=0.75) from the BLASTN website(https://blast.ncbi.nlm.nih.gov/) [89].

Phylogenetic analysis

We developed phylogenetic analysis using the concatenated coding sequences(CDS) of the chloroplast genomes from 43 species. These include 37 Lamialesspecies(*S. bowleyana*, *S. splendens*, *S. officinalis*, *S. bulleyana*, *S. digitaloides*, *S. japonica*, *S. plebeia*, *S. przewalskii*, *S. yunnanensis*, *S. miltiorrhiza*, *S. daiguii*, *S. sclarea*, *S. meiliensis*, *S. miltiorrhiza f.alba*, *S. hispanica*, *S. merjamie*, *S. petrophila*, *S. tiliifolia*, *S. chanryoenica*, *S. yangii*, *S. prattii*, *S. roborowskii*, *S. nilotica*, *R. officinalis*, *A. rugosa*, *D. heterophyllum*, *D. taliense*, *D. tanguticum*, *D. moldavica, A. forrestii, A. campylanthoides, A. ciliata, A. decumbens, A. lupulina, A. nipponensis*, *L. japonicus*, *Elsholtzia densa*), 4 species of the Verbenaceae family(*C. trichosphaera*, *C. mongholica*, *C. incana*, *C. forrestii)*,while the two species *Ligusticum chuanxiong* from the Apiaceae family and *Panax notoginseng* from the Araliaceae family were used as the outgroup. The chloroplast genome sequences were downloaded from GenBank (Additional file 1, Table S1). The shared CDS were extracted, concatenated by using PhyloSuite(v1.2.2) [90], and aligned by using MAFFT(v7.313) [91]. Phylogenetic analysis was conducted based on the maximum likelihood(ML) method implemented in IQ-TREE(v1.6.8) [92] under the TVM+F+I+G4 nucleotide substitution model. The reliability of the phylogenetic tree was assessed by bootstrap analysis with 1000 replications. Finally, the phylogenetic tree was visualized using MEGA-X [93].

Abbreviations

IGS: Intergenic sequences; cp: Chloroplast genome; CDS: Protein-coding sequence; IR: Inverted regions; ITS: Internal transcribed spacer; LSC: Large single copy; ML: Maximum likelihood; rRNA: Ribosomal RNA; SSC: Small single copy; SSR: Simple sequence repeat; tRNA: Transfer RNA; IRLC:inverted repeat-lacking clade cpDNA: Chloroplast DNA; mtDNA: Mitochondrial DNA.

Declarations

Acknowledgments

We would like to thank Dr. Mei Jiang,Dr. Haimei Chen, Mr. Haodong Chen, Rongjun Fan, Miss Xiaoying Pei, Jing Li, Yufang Ma who have provided support for data analysis.

Author' Contributions

CL conceived the study; LQW collected samples, extracted DNA for next-generation sequencing; DQ, ZJ, ZJC, SSH, and CZE assembled, validated the genome, performed data analysis, and drafted the manuscript; LQW and BW reviewed the manuscript critically. All authors have read and agreed with the contents of the manuscript.

QD: 171765300@qq.com

JZ: 2473312556@qq.com

LQW: lys832000@163.com

JCZ: joecz2021@sina.com

ZEC:cze982920577@163.com

SHS: sunsihui0311@163.com

BW: beinwang@126.com

CL: cliu6688@yahoo.com

Funding

This work was supported by funds from the Chinese Academy of Medical Sciences, Innovation Funds for Medical Sciences(CIFMS) [2021-I2M-1-022], National Science &Technology Fundamental Resources Investigation Program of China [2018FY100705], National Science Foundation [81872966], Qinghai Provincial Key Laboratory of Phytochemistry of Qinghai Tibet Plateau [2020-ZJ-Y20], Hunan technological innovation guidance project (2018SK52001). The funders were not involved in the study design, data collection, analysis, decision to publish, or manuscript preparation.

Availability of data and materials

The chloroplast genome sequence data of *S. bowleyana*、*S. splendens* and *S. officinalis* are openly available in the GenBank database with accession numbers OM617845, OM617847, and OM617846 (https://www.ncbi.nlm.nih.gov). The associated BioProject, SRA, and Bio-Sample numbers are PRJNA726222, PRJNA769231, and PRJNA769230; SAMN18926173, SAMN22106482, and SAMN22106467; SRR14415377, SRR17843445, and SRR17853381, respectively.

Ethics approval and consent to participate

All the plant materials were sampled from natural populations and no specific permission was needed to collect such samples. This study was conducted in accordance with local legislation and the Convention on the Trade in Endangered Species of Wild Fauna and Flora.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]College of pharmacy, Qinghai Minzu University, Key Laboratory of Medicinal Plant Resources of Qinghai-Tibetan Plateau in Qinghai Province, Xining, Qinghai, 810007, PR China. [2] Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing 100193, P.R.China.[3]Xiangnan University, Chenzhou, Hunan, 423000, P.R.China.[4]College of pharmacy, Heze University, Heze, Shandong Province, 274015, PR China.[5]Fresh Sky-right(Beijing) international science and technology Co., Ltd, Beijing 100187, P.R.China.

References

1. HW IC: Lamiaceae. Editorial board, Chinese Academy of Sciences. Flora of China. 17th edn. Beijing: Science Press; 1994.
2. Rattray R.D. VW, B.-E. The Botanical, Chemical and Ethnobotanical Diversity of Southern African Lamiaceae. Molecules. 2021;26:3712. https://doi.org/3710.3390/ molecules 26123712.
3. Bo Li PDC, Richard G. Olmstead, Gemma L. C. Bramley, Chun-LeiXiang, Zhong-Hui Ma, Yun-HongTan & Dian-XiangZhang. A large-scale chloroplast phylogeny of the Lamiaceae sheds new light on its subfamilial classification. Scientific Reports. 2016; 6:34343. https://doi.org/10.1038/srep34343.
4. Karpinski. TM. Essential Oils of Lamiaceae Family Plants as Antifungals. Biomolecules. 2020;10:103. https://doi.org/110.3390/biom10010103.
5. Asgar Ebadollahi MZ, and Franco Palla. Essential Oils Extracted from Different Species of the Lamiaceae Plant Family as Prospective Bioagents against Several Detrimental Pests.Molecules. 2020;25:1556.https://doi.org/10.3390/molecules250715 56.

6. Chen B, Huang C, Zhang, Y, Tang X, and Lin Y. *Salvia bowleyana* Dunn root is a novel source of salvianolic acid B and displays antitumor effects against gastric cancer cells. Oncol Lett. 2020;20:817-827. http://doi.org/10.3892/ol.2020.11611.
7. Huang XH, Zhang YF, Wang ZB, Ma XH, Fen SB, Yang NN, Zhang YH. Studies on Chemical Constituents and Tumor Cytotoxic Activity from the Root of *Salvia bowleyana*. Traditional Chinese medicinal materials. 2020;43(06):1383-1387.
8. Peng Q, Liu JX. Advances in chemical constituents and bioactivity of *Salvia genus*. Chin J Chin Mater Med. 2015;40(11):2096-2105.
9. Bai XR MY, Li YH. Investigation and research progress of the traditional application of *Salvia officinalis*. Mod Chin Med. 2019;21(2):271-278. https://doi.org/10.13313/j.issn. 1673-4 890.20180731001.
10. Afonso AF, Pereira OR, Fernandes N, Calhelha RC, Silva AMS, Ferreira ICFR, et al. Phytochemical composition and bioactive effects of *Salvia africana, Salvia officinalis* "Icterina" and *Salvia mexicana* aqueous Extracts. Molecules. 2019;24:4327. https://doi.org/10.3390/molecules24234327.
11. Green BR. Chloroplast genomes of photosynthetic eukaryotes. The Plant journal. 2019; 66(1):34-44. https://doi.org/10.1111/j.1365-313X.2011.04541.x.
12. Xiao-Ming Z, Junrui W, Li F, Sha L, Hongbo P, Lan Q, et al. Inferring the evolutionary mechanism of the chloroplast genome size by comparing whole-chloroplast genome sequences in seed plants. Sci Rep. 2019;7(1):1555. https://doi.org/10.1038/s41598 -017-01518-5.
13. Enrique Lo'pez-Juez. Plastid biogenesis, between light and shadows.J EXP BOT. 2007;58(1):11-26. https://doi.org/10.1093/jxb/erl196.
14. Glynn JM, Miyagishima S, Yoder DW, Osteryoung KW, Vitha S. Chloroplast Division. Traffic. 2007; 8:451-461. https://doi.org/10.1111/j.1600-0854.2007.00545.x
15. Palmer JD. Comparative organization of chloroplast genomes. Ann Rev Genet. 1985;19(1):325–354. https://doi.org/10.1146/annurev.ge.19.120185.001545.
16. Zhang R, Ge FF, Li HY, Chen YD, Zhao Y, Gao Y, et al. PCIR:a database of Plant Chloroplast Inverted Repeats. Database J. Biol. Databases Curation. 2019:baz127. https://doi.org/10.1093/database/baz127.
17. Henry RJ, Nock CJ, Waters DLE, Bowen SG, Rice N, Cordeiro GM, et al. Chloroplast genome sequences from total DNA for plant identification. Plant Biotechnol J. 2011; 9: 328-333. https://doi.org/10.1111/j.1467-7652.2010.00558.x.
18. Yang YC, Kung TL, Hu CY and Lin SF. Development of primer pairs from diverse chloroplast genomes for use in plant phylogenetic research. Genet Mol Res. 2015; 14(4):14857-14870. https://doi.org/10.4238/2015.November.18.51.
19. Adem M, Beyene D, Feyissa T. Recent achievements obtained by chloroplast transformation. Plant Methods. 2017;13:30. https://doi.org/10.1186/s13007-017-0179 -1.
20. Wu WG, Dong LL, Chen SL. Development direction of molecular breeding of medicinal plants. Chin J Chin Mater Med. 2020;45(11):2714-2719. https://doi.org/10.19540/ j.cnki.cjcmm.20200329.105.
21. Chiara Santos, Filipe Pereira. Identification of plant species using variable length chloroplast DNA sequences.Forensic Sci Int Genet. 2018;36:1-12. https://doi.org/ 10.1016/j.fsigen.2018.05.009.
22. Qian J, Song JY, Gao HH, Zhu YJ, Xu J, Pang XH, et al. The Complete Chloroplast Genome Sequence of the Medicinal Plant *Salvia miltiorrhiza*. PLoS ONE. 2013; 8(2):e57607. https://doi.org/10.1371/journal.pone.0057607.
23. Hu JL, Zhao M, Hou ZJ, Shang J. The complete chloroplast genome sequence of *Salvia miltiorrhiza*, a medicinal plant for preventing and treating vascular dementia. Mitochondrial DNA Part B. 2020;5(3):2460-2462. https://doi.org/10.1080/23802359. 2020.1778574.
24. Liang CL, Wang L, Lei J, Duan BZ, Ma WS, Xiao SM, et al. Comparative Analysis of the Chloroplast Genomes of Four *Salvia* Medicinal Plants. Engineering. 2019; 5:907-915. https://doi.org/10.1016/j.eng.2019.01.017.
25. Du Y, Wang YY, Xiang CL, Yang MQ. Characterization of the complete chloroplast genome of *Salvia Przewalskii* Maxim.(Lamiaceae), a substitute for Dan-Shen *Salvia miltiorrhiza* Bunge. Mitochondrial DNA Part B. 2019;4(1):981-982. https://doi.org/ 10.1080/23802359.2019.1581107.
26. Cui N, Liao BS, Liang CL, Li SF, Zhang H, Xu J, et al. Complete chloroplast genome of *Salvia plebeia*:organization,specific barcode and phylogenetic analysis. Chin J Nat Med. 2020;18(8):563-572. https://doi.org/10.1016/S1875-5364(20)30068-6.
27. Aien T, Zhao FY, Qian JF. The complete chloroplast genome sequence of the medicinal plant *Salvia yunnanensis* C. H. Wright. (Lamiaceae). Mitochondrial DNA Part B. 2019;4(2):3603-3605. https://doi.org/10.1080/23802359.2019.1677523.
28. Gao CW, Wu CH, Zhang Q, Zhao X, Wu MX, Chen RR, et al. Characterization of Chloroplast Genomes From Two *Salvia* Medicinal Plants and Gene Transfer Among Their Mitochondrial and Chloroplast Genomes. Front Genet. 2020;11:574962. https:// doi.org/10.3389/ fgene.2020.574962.
29. Cao MT, Wu JJ, Wang RH, Xu L , Qi ZC, Wei YK. The complete chloroplast genome of Russian sage *Salvia yangii* B. T. Drew (Lamiaceae). Mitochondrial DNA Part B. 2020;5(3):2590-2591. https://doi.org/10.1080/23802359.2020.1781581.
30. Ha YH, Choi KS, Choi K. Characterization of complete chloroplast genome of endemic species of Korea Peninsular, *Salvia chanryoenica*(Lamiaceae). Mitochondrial DNA Part B. 2018;3(2):992-993. https://doi.org/10.1080/23802359.2018.1495115.
31. Wang J, Feng DP, Qian J, Duan BZ, Fan M. Characterization of the complete chloroplast genome of *Salvia tiliifolia* Vahl (Lamiaceae). Mitochondrial DNA Part B. 2020;5(3):2174-2175. https://10.1080/23802359.2020.1768943.
32. Zhang XJ, Chen C, Wang R, Yao Y, Liu LX, Zhang LY. Characterization of the complete chloroplast genome of *Salvia hispanica*(Lamiaceae). Mitochondrial DNA Part B. 2020;5(2):1748-1750. https://doi.org/10.1080/23802359.2020.1749162.
33. Zhou X, Huang YB, Zhang ZC, Xu XY, Wang RH, Xu L, et al. The complete chloroplast genome of endangered Zhangjiajie sage *Salvia daiguii* Y. K. Wei & Y. B. Huang (Lamiaceae). Mitochondrial DNA Part B. 2020;5(4):3833-3834. https://doi. org/10.1080/23802359.2020.1840934.
34. Zhou Y, Zhang HR, Ping HM, Ding YN, Hu SW, Bi GY, et al. Characterization of the complete chloroplast genome of *Salvia leucantha* (Lamiaceae). Mitochondrial DNA Part B. 2021;6(12):3406-3408. https://doi. org/10.1080/23802359.2021.2000899.

35. Du Y, Wang YY, Xiang CL, Yang MQ. Characterization of the complete chloroplast genome of *Salvia trijuga* Diels (Lamiaceae). Mitochondrial DNA Part B. 2021; 6(11):3248-3249. https://doi. org/10.1080/23802359.2021.1991243.
36. Moriguchi, Y., Kang, KS., Lee, KY. et al. Genetic variation of *Picea jezoensis* populations in South Korea revealed by chloroplast,mitochondrial and nuclear DNA markers. J Plant Res. 2009;122(2):153-160. https://doi.org/10.1007/s10265-008-0210 -8.
37. Funk HT, Berg S, Krupinska K, Maier UG, Krause K. Complete DNA sequences of the plastid genomes of two parasitic flowering plant species, *Cuscuta reflexa* and *Cuscuta gronovii*. BMC Plant Biol. 2007;7:45. https://doi.org/10.1186/1471-2229-7-45.
38. McNeal JR, Kuehl JV, Boore JL, Leebens-Mack J, dePamphilis CW. Parallel loss of plastid introns and their maturase in the genus *Cuscuta*. PLoS One. 2009;4(6):e5982. https://doi.org/10.1371/journal.pone.0005982.
39. Barthet MM, Pierpont CL, Tavernier E-K. Unraveling the role of the enigmatic MatK maturase in chloroplast group IIA intron excision. Plant Direct. 2020;4(3):1–17. https://doi.org/org/10.1002/pld3.208.
40. Vogel J, Börner T, Hess WR. Comparative analysis of splicing of the complete set of chloroplast group II introns in three higher plant mutants. Nucleic Acids Res. 1999; 27(19):3866-3874. https://doi.org/org/10.1093/nar/27.19.3866.
41. Zoschke R, Nakamura M, Liere K, Sugiura M, Börner T, Schmitz-Linneweber C. An organellar maturase associates with multiple group II introns. Proc Natl Acad Sci USA. 2010;107(7):3245-3250. https://doi.org/10.1073/pnas.0909400107.
42. Leeder WM, Voskuhl S, Göringer HU. "The 2D Structure of the T. brucei Preedited RPS12 mRNA Is Not Affected by Macromolecular Crowding". Journal of Nucleic Acids. 2017; ID 6067345. https://doi.org/10.1155/2017/6067345.
43. W Weglöhner AR, Subramanian. Nucleotide sequence of a region of maize chloroplast DNA containing the 3' end of *clp*P, exon 1 of rps12 and *rpl*20 and their cotranscription. Plant Mol Biol. 1992;18(2):415-418. https://doi.org/10.1007/BF00034970.
44. Dong WP, Xu C, Li CH, Sun JH, Zuo YJ, Shi S, et al. *ycf*1, the most promising plastid DNA barcode of land plants. Sci Rep. 2015;5:8348. https://doi.org/10.1038/srep 08348.
45. Shi C, Liu Y, Huang H, Xia EH, Zhang HB, Gao LZ. Contradiction between Plastid Gene Transcription and Function Due to Complex Posttranscriptional Splicing:An Exemplary Study of *ycf*15 Function and Evolution in Angiosperms. PLoS ONE. 2013; 8(3):e59620. https://doi.org/10.1371/journal.pone.0059620.
46. Alqahtani AA, Jansen RK. The evolutionary fate of *rpl*32 and *rps*16 losses in the *Euphorbia schimperi* (Euphorbiaceae) plastome. Sci Rep. 2021;11:7466. https:// doi.org/10.1038/s41598-021-86820-z.
47. Cheatham TE, Srinivasan J, Case DA, and Kollman PA. Molecular dynamics and continuum solvent studies of the stability of polyG-polyC and polyA-polyT DNA duplexes in solution. J Biomol Struct Dyn. 1998;16(2):265-280. https://doi.org/ 10.1080/07391102.1998.10508245.
48. Niu Z, Pan J, Zhu S, Li L, Xue Q, Liu W and Ding X. Comparative Analysis of the Complete Plastomes of *Apostasia wallichii* and *Neuwiedia singapureana* (Apostasioideae) Reveals Different Evolutionary Dynamics of IR/SSC Boundary among Photosynthetic Orchids. Front Plant Sci. 2017;8:1713. https://doi.org/ 10. 3389/fpls.2017. 01713.
49. Mishra P, Kumar A, Nagireddy A, Mani DN, Shukla AK, Tiwari R, et al. DNA barcoding: an efficient tool to overcome authentication challenges in the herbal market. Plant Biotechnol J. 2016;14:8-21. https://doi.org/10.1111/pbi.12419.
50. Brodin J, Krishnamoorthy M, Athreya G, et al. A multiple-alignment based primer design algorithm for genetically highly variable DNA targets. BMC Bioinformatics, 2013;14:255. https://doi.org/10.1186/1471-2105-14-255.
51. Poczai P, Hyvönen J. Nuclear ribosomal spacer regions in plant phylogenetics: problems and prospects. Mol Biol Rep. 2010;37(4):1897-1912. https://doi.org/10.1007/ s11033-009-9630-3.
52. Yang HY, Wang LQ, Chen HM, Jiang M, Wu WW, Liu SY, et al. Phylogenetic analysis and development of molecular markers for five medicinal *Alpinia* species based on complete plastome sequences. BMC Plant Biol. 2021;21:431. https://doi.org/10.1186/ s12870-021-03204-1.
53. Conrad L, Schoch SC, Mikhail D, Carol LH, Sivakumar K, Rogneda K, Detlef L, Richard M, Kathleen O'N, Barbara R, Shobha S, Vladimir S, John PS, Lu S, Seán T, Ilene KM. NCBI Taxonomy: a comprehensive update on curation, resources and tools, Database. 2020.
54. Fisher VL. Indigenous *Salvia* Species-An Investigation of the Antimicrobial Activity, Antioxidant Activity and Chemical Composition of Leaf Extracts. 2006. http://hdl.handle.net/10539/1619.
55. Sun Q, Wang K, Yoshimura A, Doi K. Genetic differentiation for nuclear, mitochondrial and chloroplast genomes in common wild rice (*Oryza rufipogon* Griff.) and cultivated rice (*Oryza sativa* L.). Theor Appl Genet. 2002;104(8):1335-1345. https://doi.org/ 10.1007/s00122-002-0878-4.
56. Barthet MM, Hilu KW. Expression of *mat*K: Functional and evolutionary implications. Am J Bot, 2007;94:1402-1412. https://doi.org/10.3732/ajb.94.8.1402.
57. Hilu K, Liang H. The *mat*K gene:sequence variation and application in plant systematics. Am J Bot. 1997;84(6):830. https://doi.org/10.2307/2445819.
58. Barthet MM, Hilu KW. Evaluating Evolutionary Constraint on the Rapidly Evolving Gene *mat*K Using Protein Composition. J Mol Evol. 2008;66,85–97. https://doi.org/10.1007/s00239-007-9060-6.
59. Hertel S, Zoschke R, Neumann L, Qu Y, Axmann IM, Schmitz-Linneweber C. Multiple checkpoints for the expression of the chloroplast-encoded splicing factor *Mat*K. Plant Physiol. 2013;163(4):1686-1698. https://doi.org/10.1104/pp.113.227579.
60. Liu SS, Wang Z, Wang H, Su YJ, Wang T. Patterns and Rates of Plastid *rps*12 Gene Evolution Inferred in a Phylogenetic Context using Plastomic Data of Ferns. Sci Rep. 2020;10(1):9394. https://doi.org/10.1038/s41598-020-66219-y.
61. Dong WP, Liu J, Yu J, Wang L, Zhou SL. Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. PLoS One. 2012;7(4):e35071. https://doi.org/10.1371/journal.pone.0035071.

62. Jiang M, Chen HM, He SB, Wang LQ, Chen J, Liu C. Sequencing, Characterization, and Comparative Analyses of the Plastome of *Caragana rosea* var. *rosea*. Int J Mol Sci. 2018;19(5):1419. https://doi.org/10.3390/ijms19051419.

63. Cui YX, Nie LP, Sun W, Xu ZC, Wang Y, Yu J, et al. Comparative and Phylogenetic Analyses of Ginger (*Zingiber officinale*) in the Family Zingiberaceae Based on the Complete Chloroplast Genome. Plants(Basel). 2019;8(8):283. https://doi.org/10.3390/ plants8080283.

64. Amenu SG, Wei N, Wu L, Oyetola O, Hu GW, Zhou YD, et al. Phylogenomic and comparative analyses of *Coffeeae alliance* (Rubiaceae):deep insights into phylogenetic relationships and plastome evolution. BMC Plant Biol. 2022;22(1):88. https://doi.org/10.1186/ s12870-022-03480-5.

65. Bakker RT, Culham A, Gmez-Martinez R, Carvalho J, Compton J, Dawtrey R, et al. Patterns of Nucleotide Substitution in Angiosperm cpDNA *trn*L(UAA)-*trn*F(GAA) Regions. Mol Biol Evol. 2000;17(8):1146-1155. https://doi.org/10.1093/oxfordjournals. molbev.a026397.

66. Wang DY, Wu YW, Shih ACC, Wu CS, Wang YN, Chaw SM. Transfer of Chloroplast Genomic DNA to Mitochondrial Genome Occurred At Least 300 MYA. Mol Biol Evol.. 2007;24(9):2040–2048. https://doi.org/10.1093/molbev/msm133.

67. Hu GX, Takano A, Drew BT, Liu ED, Soltis DE, Soltis PS, et al. Phylogeny and staminal evolution of *Salvia* (Lamiaceae, Nepetoideae) in East Asia. Ann Bot. 2018; 122(4):649-668. https://doi.org/10.1093/aob/mcy104.

68. Ravi Gupta, Ankush Mittal, Kuldip Singh. Identifying inverted repeat structure in DNA sequences using correlation framework. European Signal Processing Conference. 2006.

69. Du Q, Jiang M, Sun SS, Wang LQ, Liu SY, Jiang CB, et al. The complete chloroplast genome sequence of *Clerodendranthus spicatus*, a medicinal plant for preventing and treating kidney diseases from Lamiaceae family. Mol Biol Rep. 2022;49(4):3073-3083. https://doi.org/10.1007/s11033-022-07135-4.

70. Boudreau E, Takahashi Y, Lemieux C, Turmel M, Rochaix JD. The chloroplast *ycf*3 and *ycf*4 open reading frames of Chlamydomonas reinhardtii are required for the accumulation of the photosystem I complex. EMBO J. 1997;16(20):6095-6104. https://doi.org/10.1093/emboj/16.20.6095.

71. Naver H, Boudreau E, Rochaix JD. Functional studies of *Ycf*3: its role in assembly of photosystem I and interactions with some of its subunits. Plant Cell. 2001; 13(12):2731-2745. https://doi.org/10.1105/tpc.010253.

72. Nellaepalli S, Ozawa SI, Kuroda H, Takahashi Y. The photosystem I assembly apparatus consisting of Ycf3-Y3IP1 and Ycf4 modules. Nat Commun. 2018;9(1):2439. https://doi.org/10.1038/s41467-018-04823-3.

73. Krech K, Ruf S, Masduki FF, Thiele W, Bednarczyk D, Albus CA, Tiller N, Hasse C, Schöttler MA, Bock R. The plastid genome-encoded *Ycf*4 protein functions as a nonessential assembly factor for photosystem I in higher plants. Plant Physiol. 2012; 159(2):579-591. https://doi.org/10.1104/pp.112.196642.

74. Vieira Ldo N, Faoro H, Fraga HP, Rogalski M, de Souza EM, de Oliveira Pedrosa F, et al. An improved protocol for intact chloroplasts and cpDNA isolation in conifers. PLoS One. 2014;9(1):e84792. https://doi.org/10.1371/journal.pone.0084792.

75. Cronn R, Liston A, Parks M, Gernandt DS, Shen R, Mockler T. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Res. 2008;36(19):e122. https://doi.org/10.1093/nar/gkn502.

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

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

78. Shi LC, Chen HM, Jiang M, Wang LQ, Wu X, Huang LF, Liu C. CPGAVAS2, an integrated plastome sequence annotator and analyzer. Nucleic Acids Res. 2019;47(W1):W65-W73. https://doi.org/10.1093/nar/gkz345.

79. Firtina C, Kim JS, Alser M, Senol Cali D, Cicek AE, Alkan C, Mutlu O. Apollo:a sequencing-technology-independent,scalable and accurate assembly polishing algorithm. Bioinformatics. 2020;36(12):3669-3679. https://doi.org/10.1093/ bioinformatics/btaa179.

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

81. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27(2):573-580. https://doi.org/10.1093/nar/27.2.573.

82. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res. 2001;29(22):4633-4642. https://doi.org/10.1093/nar/29.22.4633.

83. Amiryousefı A, Hyvönen J, Poczai P. IRscope: an online program to visualize the junction sites of chloroplast genomes. Bioinformatics. 2018;34(17):3030-3031. https://doi.org/10.1093/bioinformatics/bty220.

84. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. Nucleic Acids Res. 2004;32(Web Server issue):W273-279. https://doi.org/10.1093/nar/gkh458.

85. Brudno M, Malde S, Poliakov A, Do CB, Couronne O, Dubchak I, Batzoglou S. Glocal alignment:finding rearrangements during alignment. Bioinformatics. 2003;Suppl 1:i54-62. https://doi.org/10.1093/bioinformatics/btg1005.

86. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000;16(6):276-277. https://doi.org/10.1016/s0168- 9525(00)02024-2.

87. Mahadani AK, Awasthi S, Sanyal G, Bhattacharjee P, Pippal S. Indel-K2P: a modified Kimura 2 Parameters(K2P) model to incorporate insertion and deletion (Indel) information in phylogenetic analysis. Cyber-Physical Systems. 2021;7(1):1-13. https://doi.org/10.1080/23335777.2021.1879274.

88. Riaz T, Shehzad W, Viari A, Pompanon F, Taberlet P, Coissac E. ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. Nucleic Acids Res. 2011;39(21):e145. https://doi.org/10.1093/nar/gkr732.

89. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, and Madden TL. BLAST+: Architecture and applications. BMC Bioinform. 2009;10:421. https://doi.org/10.1186/1471-2105-10-421.

90. Zhang D, Gao F, Jakovlić I, Zou H, Zhang J, Li WX, Wang GT. PhyloSuite: An integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. Mol Ecol Resour. 2020;20(1):348-355. https://doi.org/10.1111/1755-0998.13096.
91. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772-80. https://doi.org/10.1093/molbev/mst010.
92. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE:a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32(1):268-274. https://doi.org/10.1093/molbev/msu300.
93. Hall BG. Building phylogenetic trees from molecular data with MEGA. Mol Biol Evol. 2013;30(5):1229-1235. https://doi.org/10.1093/molbev/mst012.

Figures

Figure 1

Three *Salvia* species of Lamiaceae family. *S. bowleyana*(A), *S. splendens*(B), and *S. officinalis*(C).

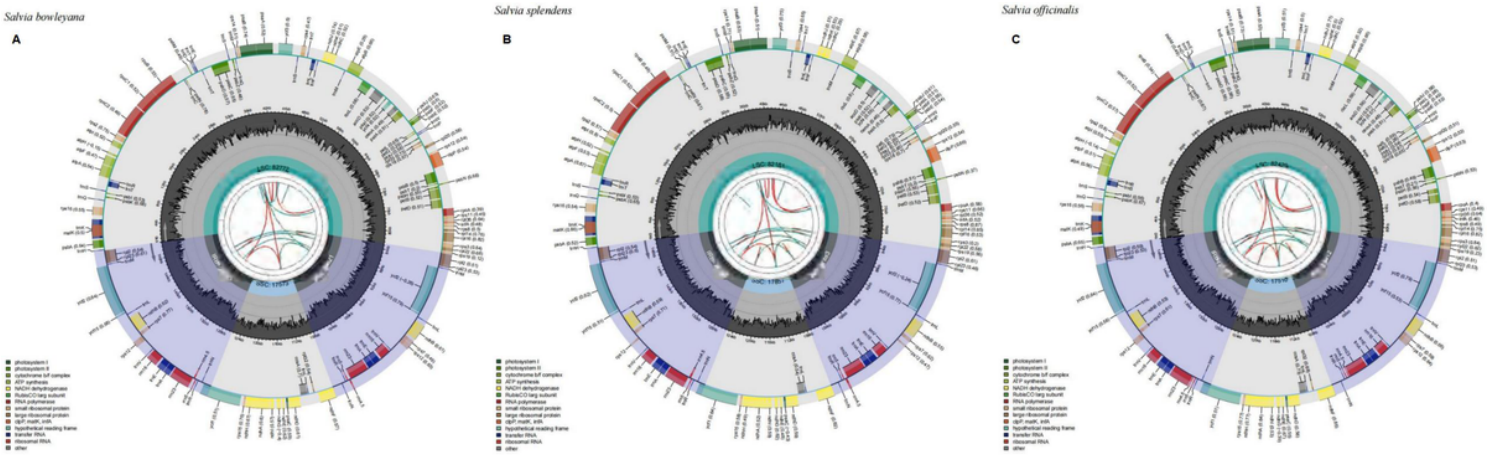

Figure 1 (A, B, C) Graphic representation of features identified in the three cp genomes of *S. bowleyana* (Figure 1A), *S. splendens* (Figure 1B) and *S. officinalis* (Figure 1C) by using CPGview-RSG (http://www.herbalgenomics.org/cpgview). The map contains seven circles. From the center going outward, the first circle shows the distributed repeats connected with red (the forward direction) and green (the reverse direction) arcs. The next circle shows the tandem repeats marked with short bars. The third circle shows the microsatellite sequences as short bars. The fourth circle shows the size of the LSC and SSC. The fifth circle shows the IRA and IRB. The sixth circle shows the GC contents along the plastome. The seventh circle shows the genes having different colors based on their functional groups.

Figure 2

See image above for figure legend

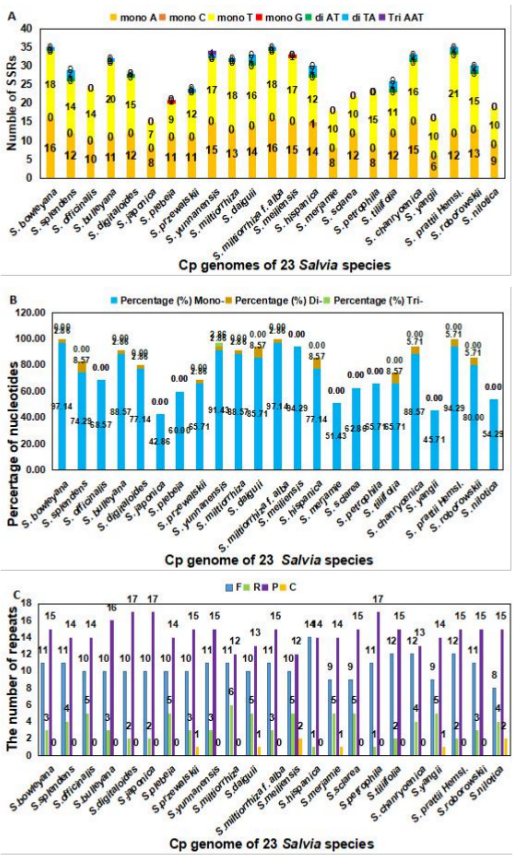

Figure 3

The repeats analysis in the 23 *Salvia* species

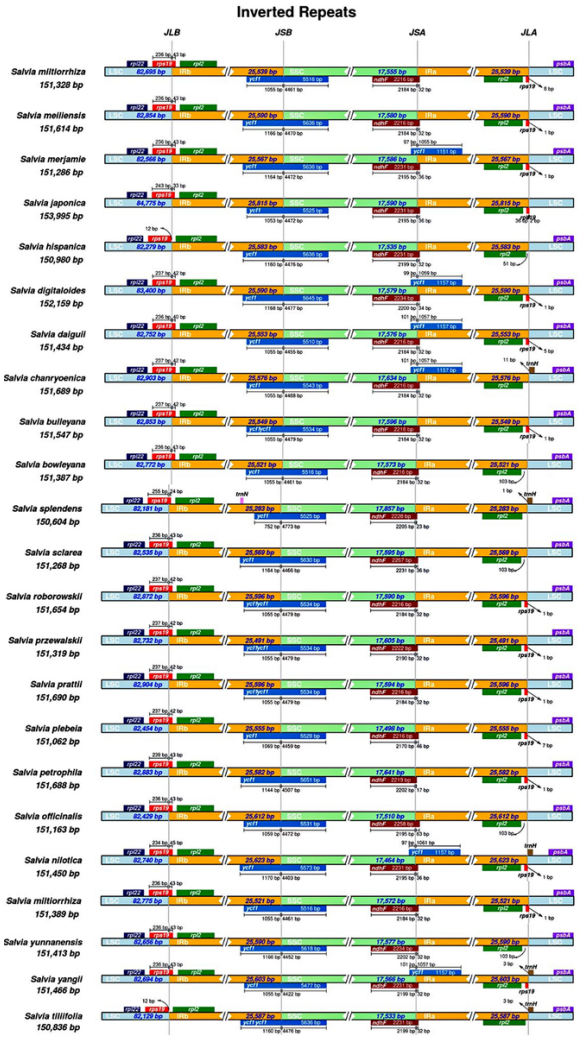

Figure 4

Comparison of the border areas among the large single-copy(LSC), small single-copy(SSC), and the inverted repeat(IR) regions in the 23 *Salvia* chloroplast genomes.

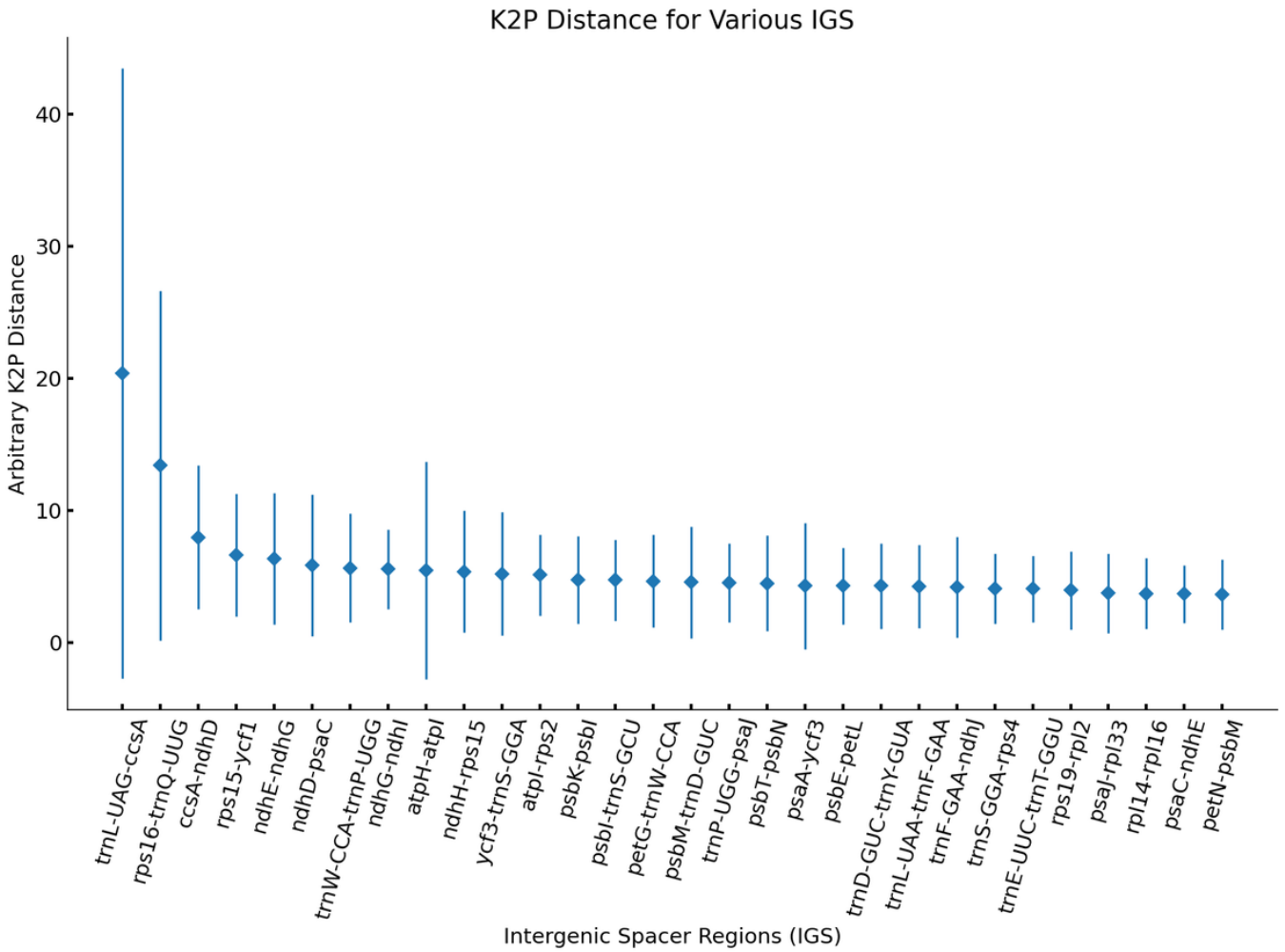

Figure 5

Average K2p distances for intergenic spacer regions in the 23 *Salvia* species from the Lamiaceae family.

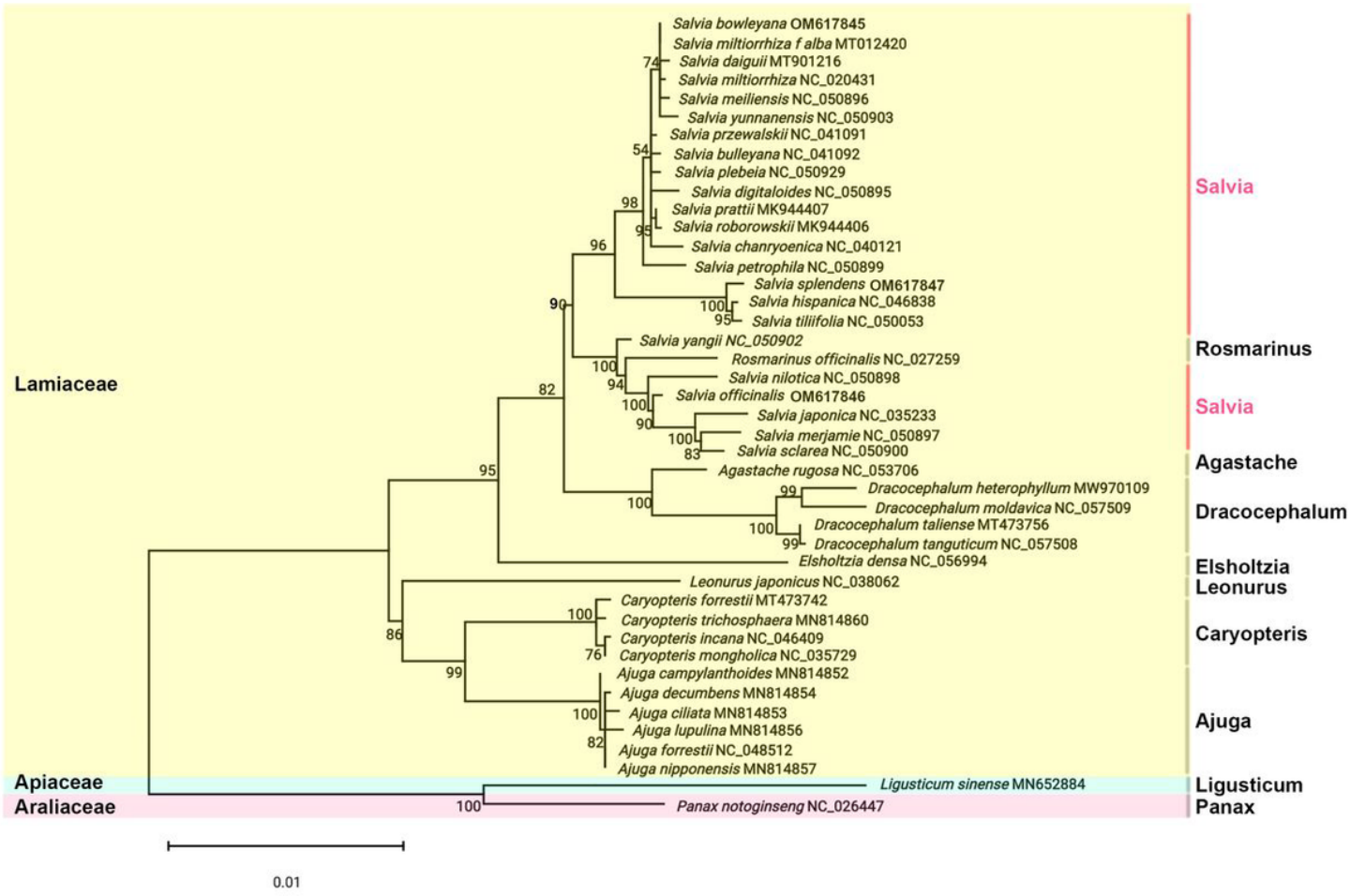

Figure 6

The phylogenetic relationships in the 43 species

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1TableS1.jpg
- Additionalfile2TableS2.jpg
- Additionalfile3TableS3.jpg
- Additionalfile4TableS4A.jpg
- Additionalfile5TableS5A.jpg
- Additionalfile6TableS6.jpg
- Additionalfile7Fig.S1.jpg
- Additionalfile8Fig.S2.jpg
- Additionalfile9Fig.S3.jpg
- Additionalfilecontent.jpg