白鼠尾草叶绿体基因组序列特征与系统进化分析

房振西[1#], 季 倩[1#], 胡佳栋[1,2], 陈万生[1,2*], 李 卿[1,2*]

(1. 上海中医药大学中药研究所, 中药资源与生物技术中心, 上海 201203; 2. 中国人民解放军海军军医大学第二附属医院药剂科, 上海 200003)

摘要: 白鼠尾草 (*Salvia apiana* Jepson) 是唇形科鼠尾草属多年生亚灌木植物, 具有悠久的药用历史。本研究采用PacBio HiFi三代测序技术对白鼠尾草叶绿体基因组进行了测序, 完成了物理图谱绘制, 对其基因组序列结构特征、密码子偏好性和重复序列进行了分析, 并与同属近缘物种进行了系统进化和叶绿体基因组比较分析。白鼠尾草叶绿体基因组序列长度为151 701 bp (GenBank登录号: OR389048), 包含典型的四分体结构, GC含量为38.06%。共注释得到132个基因, 包括87个蛋白质编码基因、37个tRNA基因和8个rRNA基因, 其中17个基因含有内含子, 18个基因为双拷贝基因。密码子偏好性分析显示, 白鼠尾草密码子偏好使用A或U结尾的密码子。白鼠尾草叶绿体重复结构分析共检测得到170个简单重复序列 (SSR) 位点和65条散在重复序列, 大部分SSR位点由A和T碱基组成。与同属21个药用植物的叶绿体全基因组系统进化分析显示, 白鼠尾草与西班牙鼠尾草 (*Salvia hispanica* Ettling. ex Willk. & Lange)、墨西哥鼠尾草 (*Salvia leucantha* Cav.) 和椴叶鼠尾草 (*Salvia tiliifolia* Vahl) 亲缘关系最近。叶绿体基因组比较分析显示, 白鼠尾草的反向重复区域 (IR) 边界存在轻微的收缩和扩张情况, 且叶绿体基因组序列中存在多处高遗传变异区。本研究建立了一种适用于利用三代测序数据组装白鼠尾草叶绿体基因组的方法, 首次对白鼠尾草叶绿体基因组进行了较为全面和深入的解析, 为白鼠尾草叶绿体基因工程、遗传多样性分析、分子育种及物种鉴定等研究提供了理论依据。

关键词: 白鼠尾草; 叶绿体基因组; 序列特征; 密码子偏好性; 系统进化

中图分类号: R931 **文献标识码**: A **文章编号**: 0513-4870(2024)05-1484-10

Characterization and phylogenetic analysis of the complete chloroplast genome of *Salvia apiana* Jepson

FANG Zhen-xi[1#], JI Qian[1#], HU Jia-dong[1,2], CHEN Wan-sheng[1,2*], LI Qing[1,2*]

(*1. Research and Development Center of Chinese Medicine Resources and Biotechnology, Institute of Chinese Materia Medica, Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China; 2. Department of Pharmacy, Second Affiliated Hospital of Naval Medical University, Shanghai 200003, China*)

Abstract: *Salvia apiana* Jepson, commonly known as white sage, is a perennial sub-shrub of the *Salvia* genus in the Lamiaceae family with a long medicinal history. In this study, the complete chloroplast genome of *S. apiana* was sequenced using PacBio HiFi third-generation sequencing technology. The physical map of the genome was constructed, and the sequence structure features, codon preference, and repetitive sequences were analyzed. Furthermore, a comparative analysis of the chloroplast genome and phylogenetic evolution with closely related species within the same genus was conducted. The chloroplast genome of *S. apiana* was found to have a length of 151 701 bp (GenBank accession number: OR389048), with a typical quadripartite structure and a GC content of

收稿日期: 2023-09-12; 修回日期: 2023-12-06.

基金项目: 国家自然科学基金资助项目 (32070327, 31770329); 国家重点研发计划项目 (2022YFC3501700).

#共同第一作者.

*通讯作者 Tel: 86-21-51322403, E-mail: chenwansheng@smmu.edu.cn; qli@smmu.edu.cn

DOI: 10.16438/j.0513-4870.2023-1056

38.06%. A total of 132 genes were annotated, including 87 protein-coding genes, 37 tRNA genes, and 8 rRNA genes. Among them, 17 genes contained introns, and 18 genes were present in duplicate copies. Codon preference analysis revealed a preference for codons ending with A or U. Analysis of repetitive structures in the *S. apiana* chloroplast genome identified 170 simple sequence repeat (SSR) sites and 65 scattered repeat sequences, with the majority of SSR sites composed of A and T. Phylogenetic analysis of the complete chloroplast genomes of 21 species within the same genus showed that *S. apiana* is most closely related to *Salvia hispanica* Ettling. ex Willk. & Lange, *Salvia leucantha* Cav., and *Salvia tiliifolia* Vahl. Comparative analysis of the chloroplast genomes revealed slight contraction and expansion of the inverted repeat (IR) boundaries in *S. apiana*, as well as multiple highly variable regions in the chloroplast genome sequence. This study establishes a method for *de novo* assembling the chloroplast genome of *S. apiana* using third-generation sequencing data and provides a comprehensive analysis of its chloroplast genome, which can serve as a theoretical basis for studies on chloroplast genetic engineering, genetic diversity analysis, molecular breeding, and species identification.

Key words: *Salvia apiana* Jepson; chloroplast genome; sequence characterization; codon preference; phylogenetic analysis

叶绿体是植物细胞内一类半自主性细胞器, 拥有独立的遗传物质——叶绿体基因组[1], 因其具有基因组结构简单和母系遗传等特点, 引起了广泛关注[2]。植物叶绿体基因组是一个约100～250 kb大小的双链环状DNA, 其中陆地植物的叶绿体基因组高度保守, 具有两个25 kb左右的反向重复区域 (inverted repeat region, IR), 以及被IR区隔离的小单拷贝区 (small single copy, SSC, 长约18～20 kb) 和大单拷贝区 (large single copy, LSC, 长约81～90 kb)[3]。大多数叶绿体基因组包含120～130个基因, 主要参与光合作用、转录和翻译等相关功能[4]。叶绿体基因组所包含的遗传信息虽然远小于核基因组, 但其结构稳定且高度保守, 在植物系统发育、物种鉴定和遗传转化等研究中具有重要作用, 已被广泛应用于药用植物研究中[5]。

白鼠尾草 (*Salvia apiana* Jepson) 隶属于唇形科 (Lamiaceae) 鼠尾草属 (*Salvia* L.), 为多年生亚灌木植物, 株高通常不到1米, 一般生长于海拔1 500米以下的干燥山坡上, 原产于美国南部和墨西哥北部地区, 我国有小部分人工种植区域[6]。白鼠尾草拥有悠久的药用历史, 其叶具有消除体味、清洁头发和防止头发变白的作用, 还能用于治疗感冒; 其种子能用于清洁眼睛; 全草燃烧后的烟雾可缓解鼻塞、多涕等症状, 用于治疗鼻窦炎和鼻炎, 其烟雾还具有驱虫灭菌和净化居所的作用[6]。丘马什人 (Chumash) 将其作为仪式用和药用植物, 常被用作烟熏剂、镇静剂、利尿剂和治疗感冒的药物[7]。现代药理学研究表明, 白鼠尾草有对抗氧化和炎症事件的能力, 还具有细胞毒性和抗菌能力, 其对肝癌细胞HepG2、宫颈癌HeLa细胞和乳腺癌细胞MCF-7均表现出较高的细胞毒性作用[8]。白鼠尾草拥有巨大的药用价值和潜力, 近年来对白鼠尾草的研究主要集中在植物分类、化学成分和药理作用等方面, 对其叶绿体基因组相关的研究尚未见报道。

本研究通过PacBio HiFi三代测序技术对白鼠尾草叶绿体基因组进行了测序、组装和注释, 并进行了序列特征、密码子偏好性、重复序列、系统进化关系以及与近缘同属植物的叶绿体基因组的比较分析, 以期为白鼠尾草叶绿体基因工程、遗传多样性分析及物种鉴定等研究提供参考。

材料与方法

植物材料与测序 从上海中医药大学药圃采集成熟期 (2022年11月) 的白鼠尾草叶片后, 委托上海凌恩生物科技有限公司进行总DNA提取以及PacBio HiFi三代测序。

叶绿体基因组序列组装及注释 将从PacBio HiFi三代测序中获得的数据经canu v2.3软件[9]进行 *de novo* 从头组装, 手动矫正后, 用IGV 2.16.1软件[10]检测组装结果的准确性。使用本地BLAST v2.12.0方法[11]分析白鼠尾草叶绿体基因组的基本结构, 统计LSC、SSC和IR区域的长度, 计算GC含量。使用在线工具CPGAVAS2 (http://47.96.249.172:16019/analyzer/annotate)[12]对组装好的白鼠尾草叶绿体基因组进行蛋白质编码区、转运RNA (tRNA) 和核糖体RNA (rRNA) 的注释分析, 并用注释校正工具Apollo v1.11.8软件[13]对注释结果进行手动校正处理。利用Organellar Genome DRAW (https://chlorobox.mpimp-golm.mpg.de/OGDraw.html) 软件[14]绘制叶绿体基因组图谱。最后将组装注释好的白鼠尾草叶绿体基因组序列提交至GenBank数据库。

密码子偏好性分析 通过PhyloSuite v1.2.2软件[15]提取白鼠尾草叶绿体基因组中所有蛋白编码序列, 使

用CodonW 1.4.2软件[16]对其叶绿体编码基因的同义密码子相对使用度 (relative synonymous codon usage, RSCU) 进行计算和分析, 采用默认参数。

重复序列分析 使用在线工具MISA (https://webblast.ipk-gatersleben.de/misa/index.php?action=1)[17]对简单重复序列 (simple sequence repeat, SSR, 也称微卫星重复序列) 进行识别和定位, 参数设置为[18]: 单核苷酸重复≥8次、二核苷酸重复≥4次、三核苷酸重复≥4次, 四、五、六核苷酸重复≥3次, 两个SSR之间的序列长度设置为0。通过在线软件REPtuter (https://bibiserv.cebitec.uni-bielefeld.de/reputer/)[19]识别和查询叶绿体基因组中的散在重复序列 (dispersed repeats), 包括正向重复 (forward repeats)、反向重复 (reverse repeats)、互补重复 (complement repeats) 以及回文重复 (palindromic repeats), 参数设定为: 最大计算重复次数 (maximum computed repeats) 5 000, 最小重复长度 (minimal repeat size) 30, Hamming距离值 (hamming distance) 3。

系统进化分析 为了揭示白鼠尾草与同属药用植物的系统进化关系, 利用国家生物信息中心叶绿体基因组信息网站 (Chloroplast Genome Information Resource, CGIR, https://ngdc.cncb.ac.cn/cgir/) 收集整理已公布的鼠尾草属药用植物叶绿体基因组共21个并下载, 以紫草科 (Boraginaceae) 玻璃苣 (*Borago officinalis* L.) 叶绿体基因组作为外类群, 用PhyloSuite v1.2.2软件[15]提取所有物种的共有蛋白编码基因并进行序列比对和优化。使用IQtree v2.0.6[20]基于最适核苷酸替换模型进行自引导1 000次重复的最大似然法 (maximum likelihood, ML) 分析, 通过MEGA v7.0.26软件[21]对ML树实现可视化。

叶绿体基因组比较分析 使用在线软件IRscope (http://genocat.tools/tools/irscope.html)[22]对白鼠尾草及其三个近缘物种——西班牙鼠尾草 (*Salvia hispanica* Ettling. ex Willk. & Lange)、墨西哥鼠尾草 (*Salvia leucantha* Cav.) 和椴叶鼠尾草 (*Salvia tiliifolia* Vahl) 进行叶绿体基因组的边界收缩和扩张分析, 以同属药用模式植物丹参 (*Salvia miltiorrhiza* Bunge)[23]作为参照。此外, 仍旧以丹参叶绿体基因组作为参考序列, 利用在线软件mVISTA (https://genome.lbl.gov/vista/mvista/submit.shtml)[24]对白鼠尾草及其近缘物种进行叶绿体基因组比较分析。

结果

1 白鼠尾草叶绿体基因组结构特征

白鼠尾草叶绿体基因组测序数据经组装矫正后, 获得总长度为151 701 bp (GenBank登录号: OR389048) 的叶绿体基因组, 测序深度在183～589×之间, 总GC含量为38.06%, 具有典型的四分体结构, 由一个LSC、一个SSC和两个IR组成 (图1)。其中两个IR区长度均为25 549 bp, GC含量为43.14%; LSC区长度为82 993 bp, GC含量为36.22%; SSC区长度为17 610 bp, GC含量为31.94%。GC含量在IR区最高, LSC区次之, SSC区含量最低。各分区的碱基组成如表1所示。

Table 1 Base composition of *S. apiana*'s chloroplast genome

Region	Length/bp	A/%	C/%	G/%	T/%	GC/%
LSC	82 993	31.16	18.57	17.65	32.62	36.22
SSC	17 610	34.23	16.81	15.13	33.83	31.94
IRa	25 549	28.49	20.77	22.37	28.37	43.14
IRb	25 549	28.37	22.37	20.77	28.49	43.14
Total	151 701	30.60	19.38	18.68	31.35	38.06

2 白鼠尾草叶绿体基因组注释结果

白鼠尾草叶绿体全基因组共注释得到132个基因, 其中蛋白质编码基因87个, tRNA基因37个, rRNA基因8个。这些基因可以根据功能分为四大类 (表2)。第一类: 光合作用相关基因, 共有46个, 包括6个ATP合酶基因、5个光合系统I基因、16个光合系统II基因、12个NADH氧化还原酶基因、6个细胞色素b/f复合体基因和1个二磷酸核酮糖羧化酶基因。第二类: 复制相关基因, 共有74个, 除tRNA和rRNA基因外, 还有29个与自我复制相关的基因, 包括4个RNA聚合酶基因、11个核糖体大亚基基因和14个核糖体小亚基基因。第三类: 6个未知功能基因。第四类: 囊膜蛋白基因 (*cemA*) 及成熟酶基因 (*matK*) 等其他6个基因。这些基因中, 有17个基因含有内含子 (除*ycf3*和*clpP*基因含有两个内含子外, 其余基因均只含一个内含子), 另有18个双拷贝基因, 均位于IR区。

3 密码子偏好性分析

白鼠尾草叶绿体基因组中共有21种氨基酸64种密码子。通过CodonW 1.4.2软件对白鼠尾草叶绿体基因组87个蛋白编码基因密码子偏好性进行分析, 共得到26 467个密码子 (图2), 其中编码亮氨酸 (Leu) 的密码子数量最多 (2 808个, 10.61%), 编码色氨酸 (Trp) 的密码子数量最少 (466个, 1.76%)。除甲硫氨酸 (Met) 和Trp只使用一个密码子AUG和UGG外, 其余氨基酸均含有2～6个同义密码子。同义密码子相对使用度 (RSCU) 分析表明 (图2), RSCU < 1的密码子共有32个, 占总密码子的一半; RSCU > 1的密码子共有30个, 占总密码子的46.88%, 其中29个密码子以A/U碱基结尾, 表现出明显的A/U偏好性, 密码子第3位也偏好以A或U结尾。RSCU > 1.5的密码子有15个, 其中精氨酸 (Arg) 偏向使用密码子AGA, RSCU值最大, 为1.85; 亮氨酸 (Leu) 的RSCU次之, 为1.84, 偏向

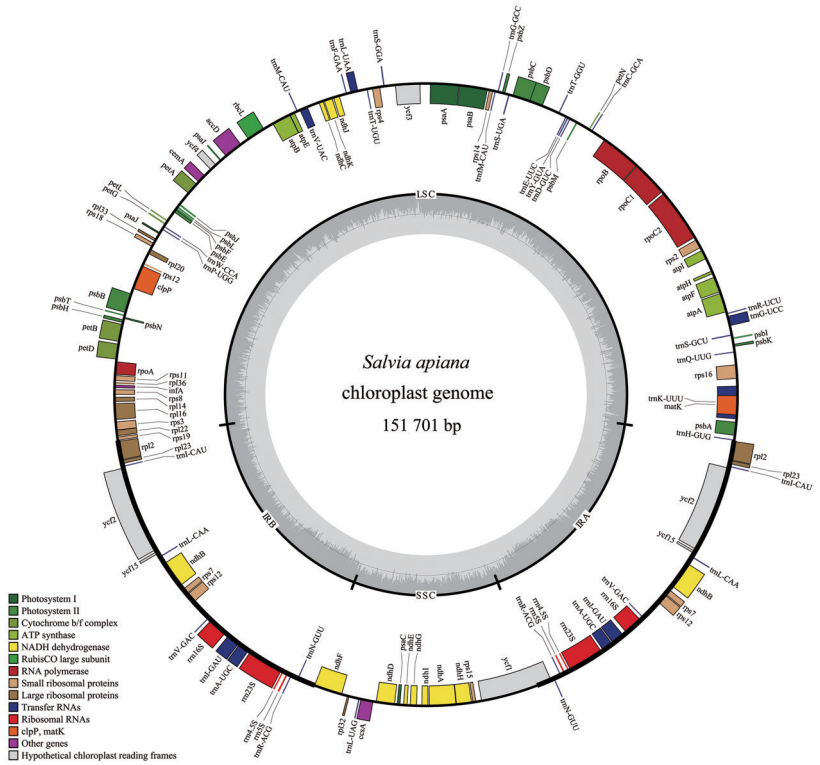

Figure 1 Chloroplast genome map of *S. apiana*. Genes are color coded by their function in the legend. Genes inside or outside the circle are transcribed in a clockwise or counter clockwise direction, respectively. The grey region of the inner circle indicates the GC content of the chloroplast genome

Table 2 Gene annotation of chloroplast genome of *S. apiana*. ×2: Copy number 2; *: Contains one intron; **: Contains two introns

Category	Group of gene	Gene name
Photosynthesis	ATP synthase	*atpA*, *atpB*, *atpE*, *atpF**, *atpH*, *atpI*
	Photosystem I	*psaA*, *psaB*, *psaC*, *psaI*, *psaJ*
	Photosystem II	*psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbJ*, *psbK*, *psbL*, *psbM*, *psbN*, *psbT*, *psbZ*, *ycf3***
	NADH-dehydrogenase	*ndhA**, *ndhB**(×2), *ndhC*, *ndhD*, *ndhE*, *ndhF*, *ndhG*, *ndhH*, *ndhI*, *ndhJ*, *ndhK*
	Cytochrome b/f complex	*petA*, *petB**, *petD**, *petG*, *petL*, *petN*
	Rubisco	*rbcL*
Self replication	DNA dependent RNA polymerase	*rpoA*, *rpoB*, *rpoC1**, *rpoC2*
	Large subunit of ribosome	*rpl14*, *rpl16**, *rpl2**(×2), *rpl20*, *rpl22*, *rpl23*(×2), *rpl32*, *rpl33*, *rpl36*
	Small subunit of ribosome	*rps11*, *rps12*(×2), *rps14*, *rps15*, *rps16**, *rps18*, *rps19*, *rps2*, *rps3*, *rps4*, *rps7*(×2), *rps8*
	rRNA	*rrn16*(×2), *rrn23*(×2), *rrn4.5*(×2), *rrn5*(×2)
	tRNA	*trnA-UGC**(×2), *trnC-GCA*, *trnD-GUC*, *trnE-UUC*, *trnF-GAA*, *trnfM-CAU*, *trnG-GCC*, *trnG-UCC**, *trnH-GUG*, *trnI-CAU*(×2), *trnI-GAU**(×2), *trnK-UUU**, *trnL-CAA*(×2), *trnL-UAA**, *trnL-UAG*, *trnM-CAU*, *trnN-GUU*(×2), *trnP-UGG*, *trnQ-UUG*, *trnR-ACG*(×2), *trnR-UCU*, *trnS-GCU*, *trnS-GGA*, *trnS-UGA*, *trnT-GGU*, *trnT-UGU*, *trnV-GAC*(×2), *trnV-UAC**, *trnW-CCA*, *trnY-GUA*
Other genes	Acetyl-CoA-carboxylase	*accD*
	c-type cytochrom synthesis gene	*ccsA*
	Envelop membrane protein	*cemA*
	Protease	*clpP***
	Translational initiation factor	*infA*
	Maturase	*matK*
Unkown	Conserved open reading frames	*ycf1*, *ycf15*(×2), *ycf2*(×2), *ycf4*

使用UUA密码子。

4 重复序列分析

白鼠尾草叶绿体基因组共检测到170个微卫星重复序列 (SSR) 位点, 其中132个为单碱基重复序列 (mononucleotide), 包括A、T和C三种重复类型; 有28个二核苷酸重复序列 (dinucleotide), 为AC、AG、

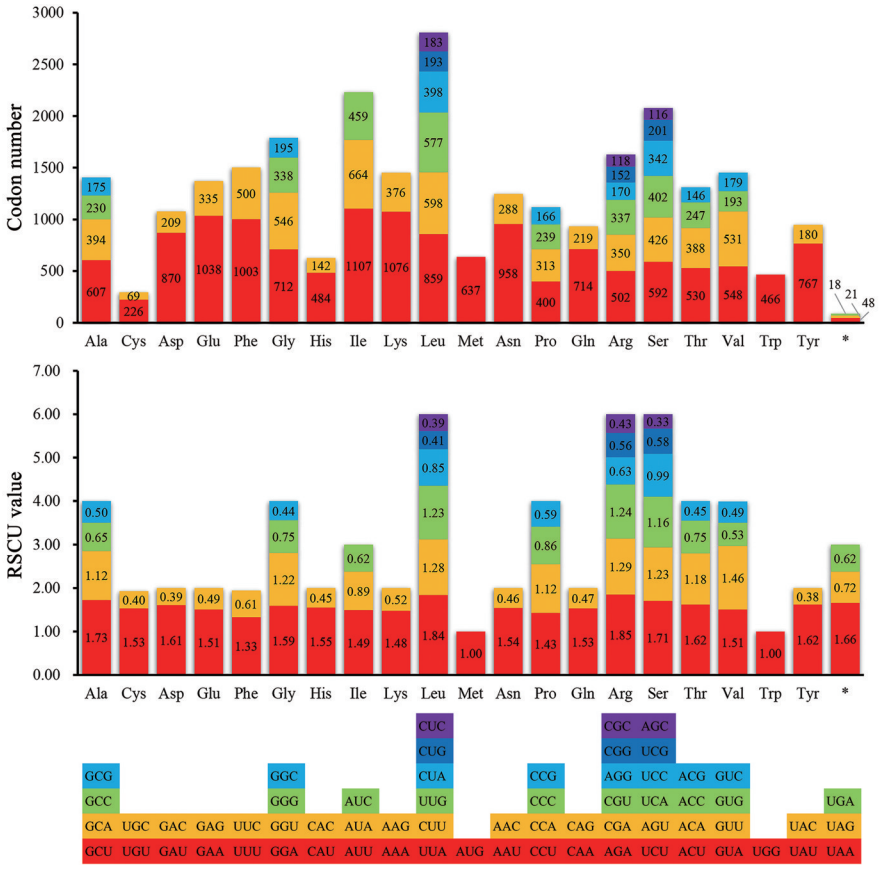

Figure 2 Codon content of 20 amino acids and stop codons in all protein-coding genes of the chloroplast genome of *S. apiana*. The histogram above each amino acid shows codon usage within *S. apiana*. Colors in the column graph reflect codons in the same colors shown below the figure. RSCU: Relative synonymous codon usage; Ala: Alanine; Cys: Cysteine; Asp: Aspartic acid; Glu: Glutamic; Phe: Phenylalanine; Gly: Glycine; His: Histidine; Ile: Isoleucine; Lys: Lysine; Leu: Leucine; Met: Methionine; Asn: Asparagine; Pro: Proline; Gln: Glutamine; Arg: Arginine; Ser: Serine; Thr: Threonine; Val: Valine; Trp: Tryptophan; Tyr: Tyrosine; *: Stop codon

AT、CA、CT、GA、TA 和 TC 八种类型; 有两个三核苷酸重复序列 (trinucleotide), 分别为 GTT 和 TTA; 还有八个四核苷酸重复序列 (tetranucleotide), 包括 AAAC、AATA、ATAA、ATTT、CTTT、GTCT、TAAA 和 TCTA 八种类型 (图 3)。白鼠尾草叶绿体基因组的 SSR 位点类型主要由 A/T 碱基组成, 表现出明显的 A/T 碱基偏好性。此外 SSR 位点在叶绿体不同区域的分布不同, LSC 区最多 (116 个, 68.24%), SSC 区次之 (30 个, 17.65%), IR 区分布最少 (24 个, 14.12%)。46 个 (27.06%) SSR 位点分布在蛋白质编码区, *rpoC2*、*ndhD*、*ycf1* 和 *ycf2* 等基因的蛋白质编码区出现多个 SSR 位点。

使用 REPuter 软件共检测到 65 个散在重复序列 (图 4), 长度范围在 30～82 bp 之间。其中正向重复 (forward repeats) 序列 34 个, 占总数的 52.31%; 回文重复 (palindromic repeats) 序列 30 个, 占总数的 46.15%; 反向重复 (reverse repeats) 序列 1 个, 占总数的 1.54%; 未出现互补重复 (complement repeats) 序列。长度为 30～39 bp 的重复序列最多, 有 45 条, 占总数的 69.23%。

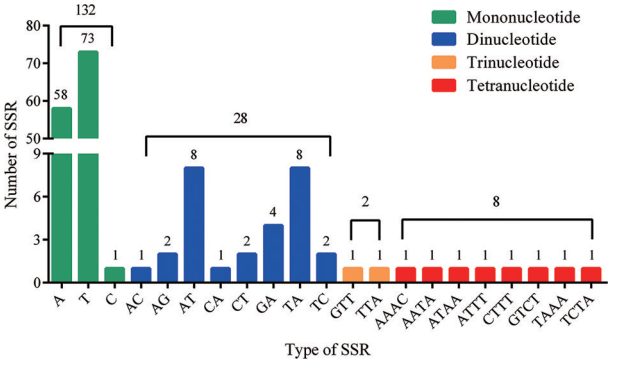

Figure 3 Types and numbers of SSRs in the chloroplast genome of *S. apiana*. SSR: Simple sequence repeat

5 系统进化分析

为了探讨白鼠尾草与同属药用植物的系统进化关系, 本研究根据国家生物信息中心叶绿体基因组信息网站收录的数据, 收集整理了所有已公布的鼠尾草属药用植物叶绿体基因组共 21 个, 连同白鼠尾草叶绿体基因组, 以紫草科玻璃苣叶绿体基因组作为外类群, 通

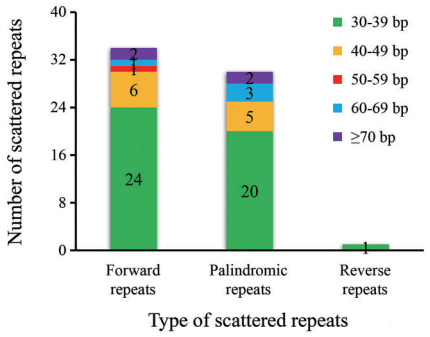

Figure 4 Types and numbers of scattered repeats in *S. apiana* chloroplast genome

过最大似然法, 以叶绿体全基因组共有基因构建分子进化树, 采用GTR+F+I+G4核苷酸替换模型。结果显示 (图5), 包括白鼠尾草在内的22种鼠尾草属药用植物分为两个主要进化枝, 进化枝I又进一步分为两个小进化枝 (Ia和Ib)。白鼠尾草与同属药用模式植物丹参同属一大枝, 但距离较远。白鼠尾草与西班牙鼠尾草、墨西哥鼠尾草和椴叶鼠尾草聚为一枝, 表明白鼠尾草与这几个物种的亲缘关系较近。

6 鼠尾草属药用植物叶绿体基因组比较分析

以药用模式植物丹参作为参考序列, 将白鼠尾草与其亲缘关系较近的同属近缘物种西班牙鼠尾草、墨西哥鼠尾草和椴叶鼠尾草叶绿体基因组进行比较分析, 它们的全长分别为151 328、151 701、150 980、151 021和150 836 bp。白鼠尾草叶绿体基因组的GC值最大, 为38.06%, 且白鼠尾草的LSC和SSC区也最长, 分别为82 993和17 610 bp (表3)。

Table 3 Chloroplast genome characteristics of five *Salvia* plants. LSC: Large single copy; SSC: Small single copy; IR: Inverted repeat region

Plant	Chloroplast genome/bp	GC/%	LSC /bp	SSC /bp	IR/bp
Salvia miltiorrhiza	151 328	38.02	82 695	17 555	25 539
Salvia apiana	151 701	38.06	82 993	17 610	25 549
Salvia hispanica	150 980	37.98	82 279	17 535	25 583
Salvia leucantha	151 021	37.99	82 262	17 537	25 611
Salvia tiliifolia	150 836	37.99	82 129	17 533	25 587

IR边界扩张和收缩分析结果显示, 五种叶绿体基因组的IR区长度在25 539～25 611 bp之间, 差异较小。这五个物种的IRa-LSC连接处均为*trnH*基因, 在西班牙鼠尾草、墨西哥鼠尾草和椴叶鼠尾草中, *trnH*基因向IRa区扩张了3 bp, 丹参和白鼠尾草的*trnH*基因均没有向IRa区扩张的趋势, 丹参的IRa-LSC连接处除*trnH*基因外还出现了*rps19*基因; IRa-SSC边界全部位于*ycf1*基因中, 除丹参的*ycf1*基因在IRa区的长度为1 056 bp外, 其余物种的*ycf1*基因在IRa区的长度均为1 161 bp; IRb-SSC边界全部位于*ndhF*基因中, 但椴叶鼠尾草的IRb-SSC边界处除了有*ndhF*基因, 还出现了*ycf1*基因, 且*ycf1*基因全部位于IRb区; 丹参的IRb-LSC边界处为*rps19*基因, 该基因向IRb区扩张了43 bp, 其余物种的IRb-LSC边界均位于*rps19*与*rpl2*基因的基因间隔区, 且*rps19*基因离边界的距离为9～12 bp不等。以上结果表明五种鼠尾草属植物的叶绿体基因组序列整体相似性较高, 存在部分边界扩张和收缩现象。

此外, 这五个鼠尾草属叶绿体基因组全序列比对

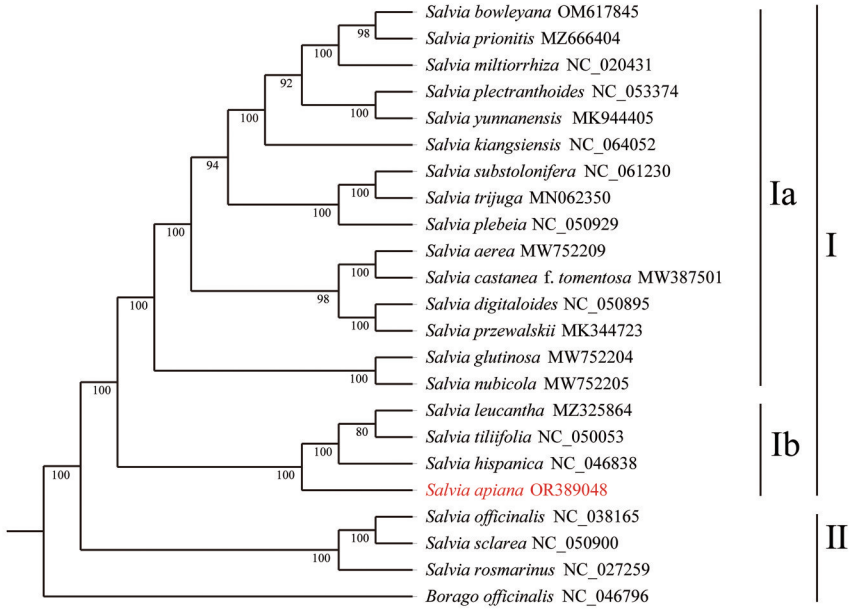

Figure 5 Phylogenetic tree based on common genes of the 23 complete chloroplast genomes

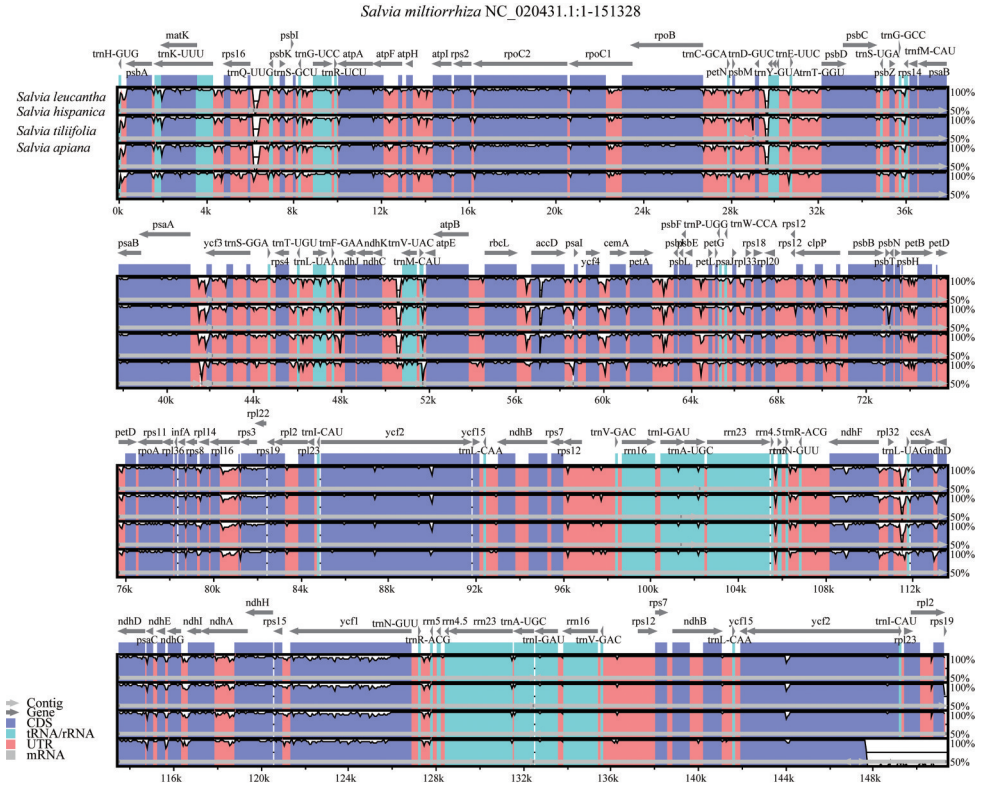

Figure 6 Comparison of chloroplast genomic structures in *Salvia* plants. Arrows indicate gene orientation; horizontal axis indicates genome size; vertical axis indicates 50%−100% similarity; coding regions, non-coding regions and tRNA or rRNA are marked on the graph with different colors

分析结果显示 (图 6), 除白鼠尾草外, 其余四种鼠尾草属植物的叶绿体基因组高度相似, 编码区相对于非编码区更加保守, 反向重复区比单拷贝区保守。各基因组位于 *trnQ-UUG-rps16*、*trnD-GUC-psbM*、*petA-psbJ*、*trnL-UAG-rpl32* 和 *psaA-ycf3* 等的基因间隔区, 以及 *accD* 和 *ycf1* 基因的内含子区的序列相似度较低, 存在不同程度的变异现象。相比于丹参叶绿体基因组, 白鼠尾草在 *ycf2*、*trnI-CAU*、*rpl23*、*rpl2* 和 *rps19* 处的差异较大。各参试物种均表现出 rRNA 和 tRNA 区域的高度保守性, 且 IR 区域相对于 LSC 和 SSC 区表现出较小的变异, 表明 IR 区在进化过程中更为保守。

讨论

叶绿体是植物细胞中由双层膜构成的含有叶绿素并能进行光合作用的一类细胞器。叶绿体基因组的研究可以追溯到 20 世纪 50 年代[25], 自第一条完整的烟草叶绿体基因组公布以来[26], 叶绿体基因组研究得到了广泛关注, 近几年以井喷方式公布了大量叶绿体基因组[27]。叶绿体基因组的获得成为了植物亲缘关系鉴定和叶绿体转化技术研究的一个重要手段[28]。

本研究基于 PacBio HiFi 三代测序技术和生物信息学分析方法, 首次对白鼠尾草叶绿体基因组进行了测序、组装和注释, 并进行了较为全面的分析, 结果显示白鼠尾草叶绿体基因组全长 151 701 bp, 具有典型的四分体结构, 与先前报道的鼠尾草属其他物种叶绿体基因组结构特征基本相似[29]。

密码子在生物遗传信息传递过程中起着重要作用, 密码子偏好性 (codon usage bias) 是生物体进化的重要特征, 对基因功能和表达存在一定影响[30]。同义密码子相对使用度 (RSCU) 作为密码子偏好性的重要指标, 用于检测编码序列内的同义密码子使用情况。当 RSCU > 1 时, 表明该密码子的使用频率较高; 若 RSCU = 1, 则表明该密码子的使用不存在偏好性; RSCU < 1, 表示该密码子使用频率较低[31]。本研究对白鼠尾草叶绿体基因组密码子偏好性进行分析, 结果显示, 白鼠尾草叶绿体基因组使用了 64 种密码子, 其中偏好密码子有 30 种, 且偏好使用 A 和 U 碱基的密码子, 与其他物种叶绿体密码子偏好性相似[32-34]。

叶绿体基因组中的简单重复序列因其含量丰富、多态性高, 且具有单亲遗传等优点已被广泛用作遗传图谱构建、靶基因校准和定位研究中的分子标记[35,36]。本研究通过对白鼠尾草叶绿体基因组的分析, 共检测得到 170 个 SSR 位点, 其中 132 个为单核苷酸重复序

列, 主要分布于LSC区。白鼠尾草叶绿体SSR位点偏向使用A或T碱基, 进一步支持了叶绿体SSR位点通常由短的polyA或polyT重复构成, 很少包含C或G串联重复的观点[37]。此外, 本研究发现SSR位点的高频CDS区, 分别为*rpoC2*、*ndhD*、*ycf1*和*ycf2*基因, 且*ycf1*基因的SSR位点最多 (10个), 提示*ycf1*基因的高变异性。SSR位点分析可为后续进一步研究白鼠尾草的分子标记、作物育种以及群体遗传分析提供参考。

在以往的系统进化学研究中发现, 采用单基因建树时, 由于有效信息位点有限, 不同基因的进化差异或选择基因序列时存在误差而得到较低分辨率的系统发育树, 无法很好地解决物种之间的进化关系[38]。本研究采用最大似然法, 基于白鼠尾草和同属21种药用植物以及一个外类群物种的75个共有蛋白编码基因进行系统进化树构建。根据Hu等[39]于2018年提出的东亚鼠尾草新的分类系统, 参与进化树构建的22个鼠尾草物种, 有20个属于东亚鼠尾草亚属 (subg. *Glutinaria* (Raf.) G.X. Hu, C.L. Xiang & B.T. Drew)。这种遗传关系反映在系统发育树中, 表现为白鼠尾草与西班牙鼠尾草、墨西哥鼠尾草和椭叶鼠尾草聚为一枝, 且这四个物种间的支持率较高, 自成一枝, 都不属于东亚鼠尾草亚属。其他物种的聚类关系, 与新的分类系统基本一致[39]。橙色鼠尾草 (*Salvia aerea* Levl.)、绒毛栗色鼠尾草 (*Salvia castanea* f. *tomentosa* Stib.)、毛地黄鼠尾草 (*Salvia digitaloides* Diels) 和甘西鼠尾草 (*Salvia przewalskii* Maxim.) 聚为一枝, 都属于东亚鼠尾草亚属的宽球苏组 (Sect. *Eurysphace*); 胶质鼠尾草 (*Salvia glutinosa* L.) 和云生丹参 (*Salvia nubicola* Wallich ex Sweet) 聚为一枝, 都是拟丹参组 (Sect. *Glutinaria*) 的成员; 佛光草 (*Salvia substolonifera* Stib.) 和三叶鼠尾草 (*Salvia trijuga* Diels) 属于截萼组 (Sect. *Substoloniferae*), 与属于荔枝草组 (Sect. *Notiosphace*) 的荔枝草 (*Salvia plebeia* R. Br.) 聚为一枝; 南丹参 (*Salvia bowleyana* Dunn)、丹参、长冠鼠尾草 (*Salvia plectranthoides* Griff.) 和云南鼠尾草 (*Salvia yunnanensis* C. H. Wright) 都属于丹参组 (Sect. *Drymosphace*), 聚在了一起, 中间却嵌入了属于须根组 (Sect. *Sobiso*) 的红根草 (*Salvia prionitis* Hance), 又与同为须根组的关公须 (*Salvia kiangsiensis* C. Y. Wu) 聚为一枝; 剩下的药鼠尾草 (*Salvia officinalis* L.)、南欧丹参 (*Salvia sclarea* L.) 和迷迭香 (*Salvia rosmarinus* Spenn.) 自成一枝, 玻璃苣为外类群物种。此系统进化结果与文献[39]不一致的地方, 可能与本研究中参与进化树构建的物种数量较少有关。

IR区域是叶绿体基因组四分结构中相对最保守的区域, IR区域边界的收缩和扩张是造成叶绿体基因组大小变化的主要原因, 其在进化中也发挥着至关重要的作用[40,41]。本研究以药用模式植物丹参为参照, 对白鼠尾草及其三个同属近缘物种进行了IR边界的扩张和收缩分析, 发现五种叶绿体基因组的IR区长度在25 539～25 611 bp之间, 差异较小, 且序列整体上高度相似, 但也存在轻微差异, 主要表现为*trnH*、*ycf1*、*ndhF*、*rps19*和*rpl2*基因的扩张和收缩。如*rps19*基因在丹参中横跨LSC-IRb边界, 而在白鼠尾草等其他四个物种中, 完全位于LSC区, 且距离IRb区9～12 bp, 表明相较于丹参, *rps19*基因在白鼠尾草等四个物种的IR区发生了收缩。一般而言, *trnH*基因在单子叶植物中位于IR区, 而在双子叶植物中位于LSC区[42], 本研究中*trnH*基因基本位于LSC区, 除了丹参和白鼠尾草的*trnH*基因位于LSC区, 西班牙鼠尾草、墨西哥鼠尾草和椭叶鼠尾草的*trnH*基因, 都向IRa区扩张了3 bp。

此外, 叶绿体基因组序列比对结果显示, 五种鼠尾草属药用植物的叶绿体基因组高度相似, 且叶绿体基因组中IR区比LSC和SSC区更保守, 编码区相对于非编码区更保守。但在*trnQ-UUG-rps16*、*trnD-GUC-psbM*、*petA-psbJ*、*trnL-UAG-rpl32*和*psaA-ycf3*等的基因间隔区, 以及*accD*和*ycf1*基因的内含子区的序列相似度较低, 有望从这些区域中开发用于鼠尾草属种间鉴定和系统发育的分子标记[43]。

叶绿体基因组是植物基因组学中的重要组成部分, 研究叶绿体基因组对于揭示叶绿体DNA的结构与起源、物种亲缘关系、植物分子标记、分子育种以及遗传转化和叶绿体基因工程等研究均具有重要意义[44]。本研究利用三代测序技术, 获得了白鼠尾草的叶绿体基因组数据, 丰富了该属的遗传资源; 同时通过生物信息学方法, 分析了其叶绿体基因组序列特征、密码子偏好性和重复序列; 基于叶绿体基因组序列对22个鼠尾草属药用植物进行了分子系统学研究; 同时通过比较基因组学分析, 提高了对白鼠尾草叶绿体基因组的认知。本研究不仅丰富了叶绿体基因组序列信息, 还为白鼠尾草叶绿体基因工程、遗传多样性分析、分子育种和物种鉴定等研究奠定了基础。

作者贡献: 房振西负责数据分析和文章撰写; 季倩参与数据分析和实验统筹; 胡佳栋负责样品采集和叶绿体基因组测序; 陈万生负责实验统筹和论文审阅; 李卿负责实验设计、叶绿体基因组组装和论文审阅。

利益冲突: 所有作者均声明不存在利益冲突。

References

[1] Corriveau JL, Coleman AW. Rapid screening method to detect potential biparental inheritance of plastid DNA and results for

over 200 angiosperm species [J]. Am J Bot, 1988, 75: 1443-1458.

[2] Fuentes P, Armarego-Marriott T, Bock R. Plastid transformation and its application in metabolic engineering [J]. Curr Opin Biotechnol, 2018, 49: 10-15.

[3] Bharadwaj R, Kumar SR, Sathishkumar R. Green biotechnology: a brief update on plastid genome engineering [J]. Adv Plant Transgenics Methods Appl, 2019, 20: 79-100.

[4] Rodríguez-Ezpeleta N, Brinkmann H, Burey SC, et al. Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes [J]. Curr Biol, 2005, 15: 1325-1330.

[5] Zhou J, Chen X, Cui Y, et al. Molecular structure and phylogenetic analyses of complete chloroplast genomes of two *Aristolochia* medicinal species [J]. Int J Mol Sci, 2017, 18: 1839.

[6] Qin X, Hu J, Fang Z, et al. Transcriptome-wide identification and expression analysis of *2-ODD* gene family in *Salvia apiana* Jepson [J]. Acta Pharm Sin (药学学报), 2022, 57: 3675-3685.

[7] Krol A, Kokotkiewicz A, Luczkiewicz M. White sage (*Salvia apiana*)-a ritual and medicinal plant of the chaparral: plant characteristics in comparison with other *Salvia* Species [J]. Planta Med, 2022, 88: 604-627.

[8] Afonso AF, Pereira OR, Fernandes ÂS, et al. The health-benefits and phytochemical profile of *Salvia apiana* and *Salvia farinacea* var. *victoria blue* decoctions [J]. Antioxidants (Basel), 2019, 8: 241.

[9] Koren S, Walenz BP, Berlin K, et al. Canu: scalable and accurate long-read assembly *via* adaptive *k*-mer weighting and repeat separation [J]. Genome Res, 2017, 27: 722-736.

[10] Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration [J]. Brief Bioinform, 2013, 14: 178-192.

[11] Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool [J]. J Mol Biol, 1990, 215: 403-410.

[12] Shi L, Chen H, Jiang M, et al. CPGAVAS2, an integrated plastome sequence annotator and analyzer [J]. Nucleic Acids Res, 2019, 47: W65-W73.

[13] Dunn NA, Unni DR, Diesh C, et al. Apollo: democratizing genome annotation [J]. PLoS Comput Biol, 2019, 15: e1006790.

[14] Greiner S, Lehwark P, Bock R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes [J]. Nucleic Acids Res, 2019, 47: W59-64.

[15] Zhang D, Gao F, Jakovlić I, et al. PhyloSuite: an integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies [J]. Mol Ecol Resour, 2020, 20: 348-355.

[16] Sharp PM, Li W. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications [J]. Nucleic Acids Res, 1987, 15: 1281-1295.

[17] Beier S, Thiel T, Münch T, et al. MISA-web: a web server for microsatellite prediction [J]. Bioinformatics, 2017, 33: 2583-2585.

[18] Qian J, Song J, Gao H, et al. The complete chloroplast genome sequence of the medicinal plant *Salvia miltiorrhiza* [J]. PLoS One, 2013, 8: e57607.

[19] Kurtz S, Choudhuri JV, Ohlebusch E, et al. REPuter: the manifold applications of repeat analysis on a genomic scale [J]. Nucleic Acids Res, 2001, 29: 4633-4642.

[20] Nguyen L, Schmidt HA, Von Haeseler A, et al. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies [J]. Mol Biol Evol, 2015, 32: 268-274.

[21] Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets [J]. Mol Biol Evol, 2016, 33: 1870-1874.

[22] Amiryousefi A, Hyvönen J, Poczai P. IRscope: an online program to visualize the junction sites of chloroplast genomes [J]. Bioinformatics, 2018, 34: 3030-3031.

[23] Song J, Luo H, Li C, et al. *Salvia miltiorrhiza* as medicinal model plant [J]. Acta Pharm Sin (药学学报), 2013, 48: 1099-1106.

[24] Frazer KA, Pachter L, Poliakov A, et al. VISTA: computational tools for comparative genomics [J]. Nucleic Acids Res, 2004, 32: W273-279.

[25] Li Q. Protective Effect of Scutellarin on Thoracic Aorta and Study of the Complete Chloroplast Genomes on *Taxus chinensis* var. *mairei* and *Cycas revoluta* (灯盏细辛血管保护作用及南方红豆杉等叶绿体基因组研究) [D]. Beijing: Peking Union Medical College, 2011.

[26] Shinozaki K, Ohme M, Tanaka M, et al. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression [J]. EMBO J, 1986, 5: 2043-2049.

[27] Zhu T, Zhang L, Cheng W, et al. Analysis of chloroplast genomes in 1 342 plants [J]. Genomics Appl Biol (基因组学与应用生物学), 2017, 36: 4323-4333.

[28] He Y, Han L, Liu Y, et al. Complete sequence analysis of chloroplast genome of *Salvia japonica* [J]. Bull Bot Res (植物研究), 2017, 37: 572-578.

[29] Liang C, Wang L, Lei J, et al. A comparative analysis of the chloroplast genomes of four *Salvia* medicinal plants [J]. Engineering, 2019, 5: 907-915.

[30] Xiong B, Wang T, Huang S, et al. Analysis of codon usage bias in xyloglucan endotransglycosylase (XET) genes [J]. Int J Mol Sci, 2023, 24: 6108.

[31] Ikemura T. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system [J]. J Mol Biol, 1981, 151: 389-409.

[32] Zhang M, Wang X, Gao J, et al. Complete chloroplast genome of *Paeonia mairei* H. Lév.: characterization and phylogeny [J]. Acta Pharm Sin (药学学报), 2020, 55: 168-176.

[33] Qiao Y, He J, Wang Y, et al. Analysis of chloroplast genome and its characteristics of medicinal plant *Sophora flavescens* [J]. Acta Pharm Sin (药学学报), 2019, 54: 2106-2112.

[34] Deng G, Wu T, Gao R, et al. Characteristics and adaptive evolution analysis of the chloroplast genome of *Gentiana rhodantha* [J]. Acta Pharm Sin (药学学报), 2022, 57: 3240-3253.

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

[36] Flannery ML, Mitchell FJG, Coyne S, et al. Plastid genome characterisation in *Brassica* and Brassicaceae using a new set of nine SSRs [J]. Theor Appl Genet, 2006, 113: 1221-1231.

[37] Kuang D, Wu H, Wang Y, et al. Complete chloroplast genome sequence of *Magnolia kwangsiensis* (Magnoliaceae): implication for DNA barcoding and population genetics [J]. Genome, 2011, 54: 663-673.

[38] Su T. Chloroplast Genome Phylogeny in *subg. Glutinaria* of Lamiaceae (唇形科东亚鼠尾草亚属叶绿体基因组系统学研究) [D]. Guiyang: Guizhou University, 2021.

[39] Hu G, Takano A, Drew BT, et al. Phylogeny and staminal evolution of *Salvia* (Lamiaceae, Nepetoideae) in East Asia [J]. Ann Bot, 2018, 122: 649-668.

[40] Yang M, Zhang X, Liu G, et al. The complete chloroplast genome sequence of date palm (*Phoenix dactylifera* L.) [J]. PLoS One, 2010, 5: e12762.

[41] Raubeson LA, Peery R, Chumley TW, et al. Comparative chloroplast genomics: analyses including new sequences from the angiosperms *Nuphar advena* and *Ranunculus macranthus* [J]. BMC Genomics, 2007, 8: 174.

[42] Li R, Ma P, Wen J, et al. Complete sequencing of five Araliaceae chloroplast genomes and the phylogenetic implications [J]. PLoS One, 2013, 8: e78568.

[43] Hu Y, Wang X, Zhang X, et al. Advancing phylogeography with chloroplast DNA markers [J]. Biodivers Sci (生物多样性), 2019, 27: 219-234.

[44] Wu L, Cui Y, Nie L, et al. The characteristics of complete chloroplast genome sequence and phylogenetic analysis of *Dendrobium moniliforme* [J]. Acta Pharm Sin (药学学报), 2020, 55: 1056-1066.