五味子叶绿体基因组结构解析与比较分析

郭豪杰¹, 刘久石¹, 罗丽¹, 魏雪苹¹, 张进², 齐耀东¹, 张本刚¹, 刘海涛^{1*}, 肖培根¹

1. 中国医学科学院北京协和医学院, 药用植物研究所, 北京 100193;

2. 新疆医科大学中医学学院, 乌鲁木齐 830011

* 联系人, E-mail: hltiu0718@126.com

收稿日期: 2017-04-10; 接受日期: 2017-05-15; 网络版发表日期: 2017-07-12

国家自然科学基金(批准号: 81373913)资助

摘要 五味子是一种重要的药用植物. 本研究通过Illumina HiSeq测序平台对五味子全基因组进行测序, 完成了五味子叶绿体基因组的组装与结构解析, 并与其他基部被子植物叶绿体基因组进行比较分析. 五味子叶绿体基因组全长146730 bp, GC含量为39.7%, 由典型的4个区域组成, 包括一个长单拷贝区、一个短单拷贝区和一对反向重复区(IR), 其长度分别为95538、18270和16461 bp. 共注释出113个基因, 包括79个蛋白质编码基因、30个tRNA基因和4个rRNA基因. 有4个蛋白质编码基因、5个tRNA基因和4个rRNA基因位于IR区. 在五味子叶绿体基因组中共检测出47个简单重复序列和40个长重复序列. 与一般被子植物叶绿体基因组相比, 五味子叶绿体基因组IR区长度大约收缩10 kb. IR区发生收缩的部分GC含量为37.4%, 远小于IR区平均GC含量(45.7%). 利用20个物种的53个共有蛋白质编码序列对五味子在被子植物中的系统发育位置进行了分析, 结果表明五味子属与八角属为姊妹分支. 对五味子叶绿体基因组序列进行分析, 为五味子科物种鉴定、遗传多样性以及系统发育研究提供了数据基础.

关键词 药用植物, 基部被子植物, IR区收缩, 系统发育分析

五味子(*Schisandra chinensis* (Turcz.) Baill.)隶属于五味子科五味子属, 其果实为常用中药材. 最早记载于《神农本草经》, 有2000多年的药用历史, 具有收敛固涩、益气生津和补肾宁心的功效^[1]. 现代药理学研究表明, 五味子具有保肝、消炎、抗肿瘤等活性^[2,3]. 五味子科多个物种在民间作为同类药材使用, 如华中五味子、翼梗五味子、红花五味子等^[4]. 为揭示其化学物质基础, 较多研究关注于五味子科植物化学成分分析, 结果表明, 五味子科不同物种的化学成分存在差异, 这可能导致它们之间的药效也有所不同^[5-7]. 因

此, 五味子科植物的物种分类与鉴定对保障临床用药的有效性具有重要意义.

五味子科植物的分类和系统发育关系一直存在争议. 学者们通常认为五味子科由五味子属和南五味子属组成, 而最近的被子植物APG(angiosperm phylogeny group)IV系统则将八角属归并入五味子科^[8-10]. 在五味子科中, 五味子属和南五味子属划分的主要依据为果序形态, 然而基于ITS和叶绿体基因片段的分子系统学研究表明, 五味子属和南五味子属物种存在交叉, 两属均不是单系^[11,12]. 此外, 该科物种划分也存在诸多争议,

引用格式: 郭豪杰, 刘久石, 罗丽, 等. 五味子叶绿体基因组结构解析与比较分析. 中国科学: 生命科学, 2017, 47: 728–739, doi: [10.1360/N052017-00127](https://doi.org/10.1360/N052017-00127)
英文版见: Guo H J, Liu J S, Luo L, et al. Complete chloroplast genome sequences of *Schisandra chinensis*: genome structure, comparative analysis, and phylogenetic relationship of basal angiosperms. Sci China Life Sci, 2017, 60, in press, doi: [10.1007/s11427-017-9098-5](https://doi.org/10.1007/s11427-017-9098-5)

物种界线不清直接导致该科一些物种鉴定困难^[13,14].

DNA条形码是利用一段标准DNA序列快速鉴定物种的技术,然而对于亲缘关系较近的物种,DNA条形码分辨率较低,目前研究表明DNA条形码技术不能解决五味子科近缘物种鉴定的问题^[15,16].被子植物叶绿体基因组长度约150 kb,变异信息远高于单条DNA条形码序列,在揭示近缘物种系统进化关系上具有较大潜力,在物种鉴定方面也具有重要价值,更被推荐为超级条形码,用于鉴定近缘物种,甚至相同物种的不同居群^[17].此外,叶绿体基因组编码区和非编码区进化速率差异大,适用于不同分类阶层的研究.随着二代高通量测序技术的发展,叶绿体基因组测序成本也不断降低,极大地促进了叶绿体基因组在居群遗传结构、系统进化以及物种鉴定等方面的研究^[18].然而,叶绿体基因组序列的数据依然严重缺乏,到目前为止,NCBI基因组数据库(<https://www.ncbi.nlm.nih.gov/genome/organelle/>)中仅有1100多条完整的叶绿体基因组.广义五味子科中,仅有少药八角(*Illicium oligandrum* Merr. & Chun)的叶绿体基因组报道,狭义五味子科中,至今未见叶绿体基因组的报道.本研究完成了五味子叶绿体基因组测序,对其结构特点进行了分析,为五味子科系统学和物种鉴定研究提供数据基础.

1 材料与方法

1.1 材料

五味子新鲜样品采自桓仁县(中国辽宁省,41°24'3"N, 124°56'37"E),经中国医学科学院药用植物研究所张本刚研究员鉴定,凭证标本保存于中国医学科学院药用植物研究所资源中心.

1.2 基因组提取和测序

利用植物DNA提取试剂盒(TIANGEN,北京)提取五味子新鲜幼嫩叶片总DNA,操作步骤参照产品说明书.利用琼脂糖凝胶电泳检测总DNA的完整性,利用NanoDrop2000微量分光光度计(Thermo Scientific,美国)检测总DNA的纯度和浓度,达到测序公司要求后,交由测序公司完成测序.利用Illumina HiSeq平台进行测序,构建长度500 bp的paired-end(PE)150文库.利用fastqc软件对测序结果进行质检,利用skewer软件、Trimmomatic软件和NGS QC tool软件去除reads上的接

头以及两端质量低于30的碱基.

1.3 叶绿体基因组组装和注释

首先,以少药八角的叶绿体基因组序列(Genbank登录号:EF380354)作为参考序列,利用BLAST++软件筛选出与参考序列相似的reads.筛选出的reads利用SOAPdenovo2软件进行组装,Kmer值设置为111^[19].第一轮组装通常产生contig数量较多,需要使用SSPACE-STANDARD-3.0软件对其进行延伸和构建scaffolding^[20].然后,以scaffolding为参考序列,再次利用BLAST++软件筛选reads用于*de novo*组装,产生的contig序列根据相邻序列两端的重叠区进行拼接,从而完成完整叶绿体基因组序列的组装.为验证序列组装的正确性,利用Genious 10.0.2软件将测序结果映射(mapping)到叶绿体基因组序列上,检测叶绿体基因组序列各碱基的覆盖度以及各contig拼接处的正确性.利用NCBI Primer-BLAST软件(<https://www.ncbi.nlm.nih.gov/tools/primer-blast/>)对反向重复区(inverted repeat region, IR)、长单拷贝区(large single copy-region, LSC)和短单拷贝区(small single copy-region, SSC)的交界处设计4对引物,用于PCR扩增,产物进行Sanger测序,以验证IR区边界的准确性(网络版附表1).利用CPGAVAS软件注释叶绿体基因组中的基因^[21].利用tRNAscan-SE 1.21软件确定tRNA基因的边界^[22].注释结果利用Apollo软件进行修订,调节编码基因起始终止位置^[23].利用sequin软件完成五味子叶绿体基因组的提交.利用OGDRAW软件画出五味子叶绿体基因组环形图谱^[24].

1.4 叶绿体基因组数据分析

利用mVISTA软件,在Shuffle-LAGAN模型下,将五味子叶绿体基因组与其他5种基部被子植物(少药八角、白睡莲(*Nymphaea alba* L.)、*Trithuria inconspicua* Cheeseman、*Amborella trichopoda* Baill.和*Nuphar advena* (Aiton) W.T. Aiton)叶绿体基因组进行比较分析^[25].利用BLAST++软件分析五味子叶绿体基因组蛋白质编码区(coding sequence, CDS)与其他基部被子植物的相似性.利用MISA软件(<http://pgrc.ipk-gatersleben.de/misa/>)检测叶绿体基因组简单重复序列(simple sequence repeats, SSR),单核苷酸重复单元不少于10个,二核苷酸和三核苷酸重复单

元不少于5个, 四核苷酸、五核苷酸和六核苷酸重复单元不少于3个. REPuter软件用于检测正向(forward, F)和回文(palindrome, P)重复序列, 参数设置: minimal repeat size 30 bp, hamming distance 3, Maximum Computed Repeats 90^[26]. Tandem repeats finder v4.04软件被用于检测串联重复(tandem, T), 参数默认^[27].

1.5 系统进化分析

基于APG IV系统选择19种被子植物完整叶绿体基因组(网络版附表2), 以裸子植物银杏(*Ginkgo biloba* L.)为外类群, 利用其53个共有基因构建最大似然(maximum-likelihood, ML)和贝叶斯(Bayesian-inference, BI)系统发育树, 用于确定五味子的系统发育位置. 序列通过ClustalW软件比对后, 利用CIPRES Science Gateway服务器(<http://www.phylo.org>)中RAxML-HPC2 7.6.3和MrBayes 3.2.2 on XSEDE工具分别构建ML和BI系统发育树^[28,29]. 利用jModelTest 2.1.4软件选择最适建树模型^[30]. 利用FigTree v1.3.1软件查看系统发育树.

2 结果与讨论

2.1 叶绿体基因组组装与结构解析

Illumina平台测序的结果, 经剪切过滤后, 共产生36085722条reads, 通过与少药八角叶绿体基因组序列进行比对, 其中约有1610000条叶绿体基因组reads, 达到约1600×的覆盖度. 组装后的叶绿体基因组经检验合格后, 提交至Genbank数据库, 登录号: KY111264.

五味子叶绿体基因组全长146730 bp, GC含量为39.7%. 其结构与大多数被子植物相似, 由典型的4个

区域组成: LSC区, SSC区和一对被分开的IR区, 长度分别为: 95538, 18270和16461 bp(表1, 图1). GC含量在五味子叶绿体基因组上分布不均, 其中IR区GC含量最高, 为45.7%, SSC区GC含量最低, 为35.0%, LSC区GC含量为38.5%. GC含量的分布不均, 可能是导致IR区保守性高于LSC和SSC区的重要因素之一^[31]. 在五味子叶绿体基因组中共注释出113个基因, 包括79个蛋白质编码基因、30个tRNA基因和4个rRNA基因, 其中有5个tRNA基因、4个蛋白质编码基因和4个rRNA基因位于两个IR区(图1, 网络版附表3). 五味子叶绿体基因组的基因顺序和组成与八角属相同, 并且与绝大部分被子植物叶绿体基因组相似, 这在一定程度上反映了被子植物叶绿体基因组具有很高的保守性^[32]. 内含子对基因选择性剪切具有重要的作用, 在五味子叶绿体基因组中, 共注释出18个含内含子的基因, 包括6个tRNA基因和12个蛋白质编码基因, 除了*ycf3*, *clpP*和*rps12*基因含有两个内含子, 其余仅含有1个内含子(表2). *trnK-UUU*基因的内含子最长, 为2482 bp. *rps12*基因为trans-spliced结构, 5'端和3'端分别位于LSC区和IR区. 值得注意的是, *ycf15*基因在大多数被子植物中存在, 但在五味子叶绿体基因组中完全消失, 相同的情况也发生在少药八角中^[33]. Shi等人^[34]研究表明, *ycf15*基因能够转录为多顺反子转录本的前体之一, 但其功能尚不清楚.

五味子叶绿体基因组CDS的总长度为72837 bp, 约占整个叶绿体基因组长度的49.5%, 由24279个密码子组成, 密码子使用率见表3. 编码亮氨酸的密码子使用率最高, 约占总使用率的10.1%, 编码半胱氨酸的密码子使用率最低, 约占总使用率的1.2%. 在CDS中, 第3个密码子的AT含量最高, 约67.9%. 这种情况也反映

表1 五味子叶绿体基因组碱基组成^{a)}

	T(U)(%)	C(%)	A(%)	G(%)	Length (bp)
Total	30.6	20.2	19.7	19.5	146730
LSC	31.3	19.7	30.1	18.9	95538
SSC	32.6	18.5	32.4	16.5	18270
IR	26.6	21.5	27.7	24.2	16461
CDS	30.7	18.3	29.9	21.1	72837
1st position	23.0	19.0	30.1	27.8	24279
2nd position	32.2	20.8	28.5	18.5	24279
3rd position	36.8	15.1	31.1	17.0	24279

a) CDS: 蛋白质编码区

图 1 五味子叶绿体基因组图谱
不同的颜色代表不同功能的基因, 灰色箭头表示基因组方向, 内圈深灰色环反应GC含量

在密码子使用率中, 同义密码子相对使用率(RSCU)大于1的密码子中, 除*trnL*-CAA对应的密码子以外, 其余均以A/U结尾。

2.2 重复序列分析

SSR是一组由1~6个碱基的重复单元串联而成的重复序列, 已经作为基因分子标记被广泛用于系统发育研究、物种鉴定以及居群遗传多样性的分析中^[35,36]。本研究从五味子叶绿体基因组中, 共检测出47个SSR, 其

中包括29个单核苷酸、6个双核苷酸、1个三核苷酸、4个四核苷酸、1个五核苷酸和6个六核苷酸SSR(表4)。非编码区包括基因间区(intergenic spacer, IGS)和内含子(intron), 绝大多数SSR分布于非编码区, 仅有7个SSR分布于*cemA*, *psbC*, *rpoC2*, *ycf1*和*ycf2*基因的CDS中。大多数被子植物叶绿体基因组中, SSR的AT含量较高^[31,37]。在五味子叶绿体基因组中也存在类似情况, 29个单核苷酸SSR中有28个为A/T组成, 其他重复单元的SSR中AT含量均较高。这种AT含量较高的情况, 在

表 2 五味子叶绿体基因组中含内含子的基因信息

基因	位置	外显子 I (bp)	内含子 I (bp)	外显子 II (bp)	内含子 II (bp)	外显子 III (bp)
<i>trnK</i> -UUU	LSC	37	2482	37		
<i>rps16</i>	LSC	41	819	220		
<i>trnG</i> -UCC	LSC	21	786	48		
<i>atpF</i>	LSC	148	733	416		
<i>rpoC1</i>	LSC	456	711	1608		
<i>ycf3</i>	LSC	127	725	230	757	153
<i>trnL</i> -UAA	LSC	35	492	50		
<i>trnV</i> -UAC	LSC	39	591	37		
<i>clpP</i>	LSC	71	801	294	556	247
<i>petB</i>	LSC	6	761	642		
<i>petD</i>	LSC	8	731	526		
<i>rpl16</i>	LSC	9	951	402		
<i>rpl2</i>	LSC	394	660	431		
<i>ndhB</i>	IR	869	609	757		
<i>trnI</i> -GAU	IR	42	935	35		
<i>trnA</i> -UGC	IR	38	798	35		
<i>ndhA</i>	SSC	556	1061	542		
<i>rps12</i>	LSC-IR	114		229		26

一定程度上与AT和GC的稳定性有关^[31]。
长重复序列为长度大于或等于30 bp的重复序列, 包括3种类型: 正向(P)、回文(P)和串联(T)重复, 可能具有促进叶绿体基因组重排的功能, 并且增加居群遗传多样性^[38]。在五味子叶绿体基因组中检测出40对长重复序列, 长度范围30~149 bp, 其中包括6个正向重复、13个回文重复和21个串联重复(表5)。绝大多数长重复序列长度小于80 bp, 仅有1对长度大于80 bp, 为149 bp。67.5%的长重复序列位于非编码区, 82.5%的长重复序列位于LSC区, 仅有5%位于SSC区。

2.3 基部被子植物比较分析

为明确五味子与其他基部被子植物的差异, 本研究将五味子叶绿体基因组与其余5条基部被子植物叶绿体基因组序列进行了比较分析。五味子叶绿体基因组基因顺序和组成与少药八角最为相似。基部被子植物叶绿体基因组长度存在明显差异, 平均长度为157727 bp, *T. inconspicua*的叶绿体基因组最长, 五味

子叶绿体基因组最短, 二者相差大约18.7 kb(网络版附 表4)。除少药八角以外, 其余基部被子植物叶绿体基因组长度变化与IR区长度变化一致。这说明, IR区长度是影响叶绿体基因组长度的重要因素之一。
以五味子叶绿体基因组为参考, 利用mVISTA软件比较分析6条基部被子植物叶绿体基因组(图2)。结果发现, 非编码区的变异明显高于编码区。IR区比LSC和SSC区保守, 这可能是由于两个IR区序列之间基因转换导致的复制校正(copy correction)^[39]。五味子叶绿体基因组与少药八角最相似, 与*T. inconspicua*差异最大。基部被子植物叶绿体基因组编码区进行比较分析, 结果显示五味子与少药八角差异最小, 为2.39%, 与*A. trichopoda*、*Nu. advena*、白睡莲、*Ny. mexicana*和*T. inconspicua*的差异分别为6.86%, 6.65%, 7.05%, 7.13%和9.67%(网络版附 表5)。差异前十的编码区为*matK*, *ycf1*, *ndhF*, *rpl22*, *rpoC2*, *accD*, *clpP*, *ccsA*, *rpoA*和*rpl20*, 全部位于LSC和SSC区, 而最保守的编码区为4个rRNA基因, 位于IR区。

表 3 五味子叶绿体基因组密码子使用率^{a)}

密码子	氨基酸	数量	RSCU	tRNA	密码子	氨基酸	数量	RSCU	tRNA
GUU	Val	511	1.47		AUU	Ile	953	1.38	
GUC	Val	171	0.49	<i>trnV</i> -GAC	AUC	Ile	458	0.66	<i>trnI</i> -GAU
GUA	Val	483	1.38	<i>trnV</i> -UAC	AUA	Ile	656	0.95	<i>trnI</i> -CAU
GUG	Val	230	0.66		CAU	His	473	1.52	
UAU	Tyr	696	1.57		CAC	His	151	0.48	<i>trnH</i> -GUG
UAC	Tyr	190	0.43	<i>trnY</i> -GUA	GGU	Gly	573	1.33	
UGG	Trp	427	1	<i>trnW</i> -CCA	GGC	Gly	180	0.42	<i>trnG</i> -GCC
ACU	Thr	490	1.57		GGA	Gly	672	1.56	<i>trnG</i> -UCC
ACA	Thr	361	1.16	<i>trnT</i> -UGU	GGG	Gly	302	0.7	
ACC	Thr	241	0.77	<i>trnT</i> -GGU	GAA	Glu	913	1.44	<i>trnE</i> -UUC
ACG	Thr	153	0.49		GAG	Glu	353	0.56	
UCU	Ser	513	1.65		CAA	Gln	609	1.46	<i>trnQ</i> -UUG
UCA	Ser	398	1.28	<i>trnS</i> -UGA	CAG	Gln	224	0.54	
AGU	Ser	360	1.16		UGU	Cys	213	1.49	
UCC	Ser	303	0.97	<i>trnS</i> -GGA	UGC	Cys	73	0.51	<i>trnC</i> -GCA
UCG	Ser	174	0.56		GAU	Asp	764	1.57	
AGC	Ser	121	0.39	<i>trnS</i> -GCU	GAC	Asp	210	0.43	<i>trnD</i> -GUC
CCU	Pro	394	1.52		AAU	Asn	819	1.51	
CCA	Pro	303	1.17	<i>trnP</i> -UGG	AAC	Asn	268	0.49	<i>trnN</i> -GUU
CCC	Pro	228	0.88		CGU	Arg	324	1.26	<i>trnR</i> -ACG
CCG	Pro	115	0.44		CGC	Arg	110	0.43	
UUU	Phe	763	1.17		CGA	Arg	326	1.27	
UUC	Phe	542	0.83	<i>trnF</i> -GAA	CGG	Arg	119	0.46	
AUG	Met	588	1	<i>trnI</i> /M-CAU, <i>trnM</i> -CAU	AGA	Arg	480	1.86	<i>trnR</i> -UCU
AAA	Lys	835	1.41	<i>trnK</i> -UUU	AGG	Arg	186	0.72	
AAG	Lys	350	0.59		GCU	Ala	609	1.76	
UUA	Leu	695	1.7	<i>trnL</i> -UAA	GCC	Ala	237	0.68	
UUG	Leu	526	1.28	<i>trnL</i> -CAA	GCA	Ala	405	1.17	<i>trnA</i> -UGC
CUU	Leu	489	1.19		GCG	Ala	137	0.39	
CUA	Leu	347	0.85	<i>trnL</i> -UAG	UAA	*	33	1.19	
CUG	Leu	208	0.51		UAG	*	24	0.87	
CUC	Leu	194	0.47		UGA	*	26	0.94	

a) RSCU: 同义密码子相对使用率

2.4 IR区收缩与扩张分析

IR区的收缩与扩张可能是被子植物叶绿体基因组长度多样性的主要原因^[40]. 比较五味子、少药八角、白睡莲、*Nu. advena*、*T. inconspicua*和*A. trichopoda*叶

绿体基因组IR区与LSC和SSC区的边界, 结果发现, 基部被子植物叶绿体基因组IR区收缩与扩张十分剧烈(图3). 白睡莲、*Ny. mexicana*、*T. inconspicua*和*A. trichopoda*的LSC-IRa交界位于*rps19*与*rpl2*基因之间.

图2 基部被子植物叶绿体基因组序列比较分析

每个基因上的灰色箭头表示该基因的方向; Y轴表示50%~100%的相似性; X轴表示叶绿体基因组坐标; 不同颜色的波形图表示叶绿体基因组不同区域

*Nu. advena*的LSC-IRa交界处于*rps19*与*trnH*基因之间,造成这种情况的原因是,*Nu. advena*的IR区在IRb-LSC交界处发生了扩张,使得*trnH*基因转移到IRb区。除*Nu. advena*外,其余物种的*trnH*基因都位于LSC区。五味子和少药八角的IR区在LSC-IRa交界处发生了明显的收缩,使得该交界位于*trnL-CAA*和*ndhB*基因之间。除*T. inconspicua*外,其余物种的叶绿体基因组SSC-IRb交界都位于*ycf1*基因中,从而导致这些物种在IRa区产生一个*ycf1*假基因。*T. inconspicua*在SSC-IRb处发生了明显

的扩张,使得该交界处于*ndhD*基因中,从而在IRa区产生一个*ndhD*假基因。

与一般的被子植物叶绿体基因组相比(本研究以葡萄(*Vitis vinifera* L.)为一般被子植物代表),五味子IR区发生明显收缩现象,*trnL-CAA*,*ycf15*,*ycf2*,*trnL-CAU*和*rpl23*基因转移到LSC区。类似的情况也发生在少药八角和芫荽(*Coriandrum sativum* L.)叶绿体基因组中^[33,41]。IR区扩张已有较多的研究报道,并对其扩张机制进行了讨论,主要观点认为,轻微IR扩张可能由基因转移导

表 4 五味子叶绿体基因组简单重复序列(SSR)信息^{a)}

编码	类型	SSR	长度(bp)	起始	终止	位置
1	p1	(A)10	10	464	473	IGS
2	p1	(A)15	15	4889	4903	IGS
3	p1	(T)10	10	9264	9273	IGS
4	c	(T)10attccagtaattccagtaagtagtaagt(A)10	48	9878	9925	IGS
5	p1	(A)11	11	10228	10238	IGS
6	p4	(TTAT)3	12	13876	13887	intron
7	p2	(AT)10	20	15772	15791	IGS
8	p1	(T)10	10	20242	20251	rpoC2
9	p5	(AGAAT)3	15	29834	29848	IGS
10	p1	(T)13	13	31537	31549	IGS
11	p2	(TA)7	14	34258	34271	IGS
12	p1	(T)10	10	34452	34461	IGS
13	p1	(G)10	10	36658	36667	psbC
14	p2	(AT)5	10	38546	38555	IGS
15	p4	(AAGA)3	12	45143	45154	intron
16	p1	(A)10	10	46479	46488	intron
17	p6	(TTTGAA)9	54	47486	47539	IGS
18	p6	(CATAAG)3	18	48957	48974	IGS
19	p2	(AT)6	12	52809	52820	IGS
20	p1	(A)10	10	53253	53262	IGS
21	p1	(A)10	10	56738	56747	IGS
22	p1	(T)12	12	57090	57101	IGS
23	p1	(A)10	10	57235	57244	IGS
24	p2	(TA)5	10	61970	61979	IGS
25	p2	(TC)5	10	63745	63754	cemA
26	p4	(AATG)3	12	64355	64366	cemA
27	p1	(A)12	12	69379	69390	IGS
28	p1	(A)17	17	70705	70721	IGS
29	p6	(TATTAC)3	18	74480	74497	intron
30	p1	(T)12	12	80050	80061	IGS
31	p1	(A)10	10	81922	81931	IGS
32	p6	(TTAAAG)3	18	82761	82778	IGS
33	p1	(T)13	13	83321	83333	IGS
34	c	(TTCT)3ctcaccctccatttatctacatcctcta(T)10	52	84463	84514	intron
35	p1	(A)11	11	91682	91692	ycf2
36	p6	(CAAGAA)3	18	94724	94741	ycf2
37	p1	(A)10	10	101854	101863	IGS
38	p1	(A)12	12	109553	109564	IGS
39	p1	(T)10	10	114410	114419	IGS
40	p1	(T)10	10	116418	116427	IGS
41	c	(T)11gttcctttctttccacagataggattttcccgatcagg(AAT)5	66	125681	125746	IGS
42	p6	(AAAAGA)3	18	129434	129451	ycf1
43	p1	(T)12	12	133023	133034	IGS
44	p1	(T)10	10	140724	140733	IGS

a) p: 单个的SSR类型; c: 复合的SSR类型; IGS: 基因间区; intron: 内含子

 

致,而较大的IR扩张可能是通过双链断裂修复(double-strand break repair, DSBR)机制实现的^[42,43].然而,IR区收缩的报道较少,目前主要认为IR区收缩机制同样为DSBR机制^[41].有研究表明,重复序列或polyA结构与tRNA之间的基因重组是IR区长度变化的原因之一^[44].比较五味子、少药八角和芫荽叶绿体基因组IR区收缩部分与IR区平均GC含量发现,收缩部分的GC含量远小于IR区(网络版附表6).因此推测,GC含量可能与IR区收缩存在一定联系.

2.5 系统进化分析

叶绿体基因组对了解植物进化具有重要价值^[45].本研究利用20个物种的53个共用蛋白质编码基因进行了系统进化分析.比对后序列长33339 bp,其中有12614个变异位点和6740个最简信息位点.通过构建ML树和BI树,来确定五味子科在基部被子植物中的系统发育位置(网络版附 图1和2).

ML树和BI树的拓扑结构一致,共有5个分支,标记为Clade A-E(图4).Clade A由两个亚支组成,Clade A1和A2,Clade A1为双子叶植物分支,Clade A2为单子叶植物分支.Clade B和C是两个单系分支,分别为金粟兰目和木兰类植物.Clade D和E为基部被子植物,Clade D为木兰藤目植物,Clade E由E1和E2两个亚支组成,分

别为睡莲目和无油樟目.五味子和少药八角为姊妹类群,共同构成Clade D,属于木兰藤目,与APG IV分类系统相似^[10].此外,有研究认为金粟兰目是木兰类植物的姊妹分支^[9],然而本研究结果显示,金粟兰目比木兰类植物进化,并且单独构成一个分支(Clade B),与APG IV分类系统相似^[10,46].无油樟目一直被认为是最基部的被子植物,近年来研究显示,睡莲目和无油樟目为姊妹类群,共同构成被子植物最基部的分支^[9,10,47-50].本研究的结果支持睡莲目和无油樟目共同构成最基部被子植物(Clade E).

3 结论

本研究完成了五味子叶绿体基因组测序、组装与注释,并对其结构、GC含量、密码子使用率、重复序列等进行了分析.结果显示,五味子叶绿体基因组结构和基因组成与大多数被子植物类似,但*ycf15*基因缺失.与大多数被子植物相比,五味子叶绿体基因组的IR区发生了约10 kb的收缩.通过叶绿体基因组系统进化分析,五味子的系统发育位置得到确认.五味子作为传统药用植物,具有重要药用和经济价值,其叶绿体基因组的测序与分析,为五味子科系统进化分析和物种鉴定提供了数据基础,并为解决近缘种进化关系和物种鉴定等提供了新的方法.

参考文献

- 1 国家药典委员会. 中华人民共和国药典. 北京: 中国医药科技出版社. 2015. 66-67; 244; 362-363
- 2 Szopa A, Ekiert R, Ekiert H. Current knowledge of *Schisandra chinensis* (Turcz.) Baill. (Chinese magnolia vine) as a medicinal plant species: a review on the bioactive components, pharmacological properties, analytical and biotechnological studies. *Phytochem Rev*, 2017, 16: 195-218
- 3 Panossian A, Wikman G. Pharmacology of *Schisandra chinensis* Bail.: an overview of Russian research and uses in medicine. *J Ethnopharmacol*, 2008, 118: 183-212
- 4 国家中医药管理局. 中华本草. 北京: 科学出版社, 1996
- 5 Shi Y M, Xiao W L, Pu J X, et al. Triterpenoids from the Schisandraceae family: an update. *Nat Prod Rep*, 2015, 32: 367-410
- 6 Liu H, Liu J, Zhang J, et al. Chemical constituents in plants of genus *Kadsura* Kaempf. ex Juss. *Chin Herbal Meds*, 2014, 6: 172-197
- 7 Liu J, Qi Y, Lai H, et al. Genus *Kadsura*, a good source with considerable characteristic chemical constituents and potential bioactivities. *Phytomedicine*, 2014, 21: 1092-1097
- 8 Smith A C. The families Illiciaceae and Schisandraceae. *Sargentia*, 1947, 7: 1-224
- 9 Bremer B, Bremer K, Chase M, et al. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Bot J Linn Soc*, 2009, 161: 105-121
- 10 The Angiosperm Phylogeny Group. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Bot J Linn Soc*, 2016, 181: 1-20
- 11 Liu Z, Hao G, Luo Y, et al. Phylogeny and androecial evolution in Schisandraceae, inferred from sequences of nuclear ribosomal DNA ITS and chloroplast DNA *trnL-F* regions. *Int J Plant Sci*, 2006, 167: 539-550

- 12 Wang Y H, Zhang S Z, Gao J P, et al. Phylogeny of the Schisandraceae based on cpDNA *mat-K* and *rpL16* intron data. *Chem Biodivers*, 2006, 3: 359–369
- 13 Saunders R M K. Monograph of *Schisandra* (Schisandraceae). *Syst Bot Monographs*, 2000, 58: 1–149
- 14 高石曼, 郭豪杰, 齐耀东, 等. 滇鸡血藤药材基原植物的探讨. *中药材*, 2015, 38: 2644–2650
- 15 Hebert P D N, Cywinska A, Ball S L, et al. Biological identifications through DNA barcodes. *Proc R Soc B-Biol Sci*, 2003, 270: 313–321
- 16 Zhang J, Chen M, Dong X, et al. Evaluation of four commonly used DNA barcoding loci for Chinese medicinal plants of the family Schisandraceae. *PLoS ONE*, 2015, 10: e0125574
- 17 Li X, Yang Y, Henry R J, et al. Plant DNA barcoding: from gene to genome. *Biol Rev*, 2015, 90: 157–166
- 18 Tangphatsornruang S, Sangsakru D, Chanprasert J, et al. The chloroplast genome sequence of mungbean (*Vigna radiata*) determined by high-throughput pyrosequencing: structural organization and phylogenetic relationships. *DNA Res*, 2010, 17: 11–22
- 19 Luo R, Liu B, Xie Y, et al. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *GigaSci*, 2012, 1: 18
- 20 Boetzer M, Henkel C V, Jansen H J, et al. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*, 2011, 27: 578–579
- 21 Liu C, Shi L, Zhu Y, et al. CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. *BMC Genomics*, 2012, 13: 715
- 22 Schattner P, Brooks A N, Lowe T M. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res*, 2005, 33: W686–W689
- 23 Lee E, Harris N, Gibson M, et al. Apollo: a community resource for genome annotation editing. *Bioinformatics*, 2009, 25: 1836–1837
- 24 Lohse M, Drechsel O, Kahlau S, et al. OrganellarGenomeDRAW—a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Res*, 2013, 41: W575–W581
- 25 Frazer K A, Pachter L, Poliakov A, et al. VISTA: computational tools for comparative genomics. *Nucleic Acids Res*, 2004, 32: W273–W279
- 26 Kurtz S, Choudhuri J V, Ohlebusch E, et al. REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res*, 2001, 29: 4633–4642
- 27 Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*, 1999, 27: 573–580
- 28 Ronquist F, Huelsenbeck J P. MrBayes 3: bayesian phylogenetic inference under mixed models. *Bioinformatics*, 2003, 19: 1572–1574
- 29 Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 2006, 22: 2688–2690
- 30 Posada D. jModelTest: phylogenetic model averaging. *Mol Biol Evol*, 2008, 25: 1253–1256
- 31 Yang Y, Dang Y, Yuan Y, et al. Complete chloroplast genome sequence of poisonous and medicinal plant *Datura stramonium*: organizations and implications for genetic engineering. *PLoS ONE*, 2014, 9: e110656
- 32 Wicke S, Schneeweiss G M, dePamphilis C W, et al. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Mol Biol*, 2011, 76: 273–297
- 33 Hansen D R, Dastidar S G, Cai Z, et al. Phylogenetic and evolutionary implications of complete chloroplast genome sequences of four early-diverging angiosperms: *Buxus* (Buxaceae), *Chloranthus* (Chloranthaceae), *Dioscorea* (Dioscoreaceae), and *Illicium* (Schisandraceae). *Mol Phylogenet Evol*, 2007, 45: 547–563
- 34 Shi C, Liu Y, Huang H, et al. Contradiction between plastid gene transcription and function due to complex posttranscriptional splicing: an exemplary study of *ycf15* function and evolution in angiosperms. *PLoS ONE*, 2013, 8: e59620
- 35 Dashnow H, Tan S, Das D, et al. Genotyping microsatellites in next-generation sequencing data. *BMC Bioinformatics*, 2015, 16: A5
- 36 Chmielewski M, Meyza K, Chybicki I, et al. Chloroplast microsatellites as a tool for phylogeographic studies: the case of white oaks in Poland. *iForest*, 2015, 8: 765–771
- 37 Raveendar S, Na Y W, Lee J R, et al. The complete chloroplast genome of *Capsicum annuum* var. *glabriusculum* using Illumina sequencing. *Molecules*, 2015, 20: 13080–13088
- 38 Qian J, Song J, Gao H, et al. The complete chloroplast genome sequence of the medicinal plant *Salvia miltiorrhiza*. *PLoS ONE*, 2013, 8: e57607
- 39 Khakhlova O, Bock R. Elimination of deleterious mutations in plastid genomes by gene conversion. *Plant J*, 2006, 46: 85–94
- 40 Kim K J, Lee H L. Complete chloroplast genome sequences from Korean ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. *DNA Res*, 2004, 11: 247–261
- 41 Peery R M. Understanding angiosperm genome interactions and evolution: insights from sacred lotus (*Nelumbo nucifera*) and the carrot family (Apiaceae). Doctor Dissertation. Illinois: University of Illinois at Urbana-Champaign, 2015
- 42 Goulding S E, Wolfe K H, Olmstead R G, et al. Ebb and flow of the chloroplast inverted repeat. *Molec Gen Genet*, 1996, 252: 195–206
- 43 Wang R J, Cheng C L, Chang C C, et al. Dynamics and evolution of the inverted repeat-large single copy junctions in the chloroplast genomes

- of monocots. [BMC Evol Biol](#), 2008, 8: 36
- 44 Plunkett G M, Downie S R. Expansion and contraction of the chloroplast inverted repeat in Apiaceae subfamily Apioideae. [Syst Bot](#), 2000, 25: 648–667
- 45 Wu M, Lan S, Cai B, et al. The complete chloroplast genome of *Guadua angustifolia* and comparative analyses of neotropical-paleotropical bamboos. [PLoS ONE](#), 2015, 10: e0143792
- 46 Qiu Y L, Lee J, Bernasconi-Quadroni F, et al. The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. [Nature](#), 1999, 402: 404–407
- 47 Soltis D E, Smith S A, Cellinese N, et al. Angiosperm phylogeny: 17 genes, 640 taxa. [Am J Bot](#), 2011, 98: 704–730
- 48 Goremykin V V, Nikiforova S V, Biggs P J, et al. The evolutionary root of flowering plants. [Syst Biol](#), 2013, 62: 50–61
- 49 Finet C, Timme R E, Delwiche C F, et al. Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. [Curr Biol](#), 2010, 20: 2217–2222
- 50 Qiu Y L, Li L, Wang B, et al. Angiosperm phylogeny inferred from sequences of four mitochondrial genes. [J Syst Evol](#), 2010, 48: 391–425