药食两用藠头叶绿体基因组解析、比较基因组学及系统发育研究

杨俏俏, 姜 梅, 王立强, 陈海梅, 刘 昶*, 黄林芳*

(中国医学科学院、北京协和医学院药用植物研究所, 国家中医药管理局中药资源保护重点研究室, 北京 100193)

摘要: 藠头 (*Allium chinense*) 为百合科葱属常用药食两用植物。为探究部分葱属植物系统发育关系不明确及该属物种鉴别通用DNA条形码匮乏的问题, 本研究应用高通量测序技术对藠头叶绿体全基因组进行测序、组装、注释及系统进化研究。结果显示藠头叶绿体基因组为152 525 bp, 呈典型的四分状结构, 共编码116个基因, 其中蛋白质编码基因81个, 转运RNA (transfer RNA) 基因31个和核糖体RNA (ribosome RNA) 基因4个。对6个葱属植物叶绿体基因组比较分析发现了7个变异较大的区间, 包括基因编码区 *ndhA* 和 *ycf1*, 以及非编码区 *rps16-trnQ*、*trnT-trnF*、*ndhF-rpl32*、*rpl32-trnL* 和 *rpl16-rps3*。利用葱属和非葱属53个物种的58个共有蛋白序列构建了进化树, 发现除百合属、重楼属外其余各属物种所在分枝的支持率都达到66%～100%, 有效地解决该类植物的系统进化与分类问题。此外, 利用EcoPrimer软件成功发现了7个可用于葱属物种鉴定的候选DNA条形码序列并设计了引物。本研究首次获得了藠头叶绿体基因组序列, 明确葱属物种间的亲缘关系, 发现了一系列葱属物种特异的DNA条形码序列, 为深入研究葱属植物的系统进化、分类及物种鉴定提供科学依据。

关键词: 藠头; 叶绿体基因组; 分子标记; 系统发育; DNA条形码

中图分类号: R931 **文献标识码**: A **文章编号**: 0513-4870(2019)01-0173-09

Complete chloroplast genome of *Allium chinense*: comparative genomic and phylogenetic analysis

YANG Qiao-qiao, JIANG Mei, WANG Li-qiang, CHEN Hai-mei, LIU Chang*, HUANG Lin-fang*

(*Key Research Laboratory of Traditional Chinese Medicine Resources Protection, State Administration of Traditional Chinese Medicine of the People's Republic of China, Institute of Medicinal Plants, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100193, China*)

Abstract: *Allium chinense* belongs to the genus Alliums of the lily family. It can be used both as medicine and food. To date, the phylogenetic relationship of *Allium* species have not be resolved completely. Furthermore, there has been a lack of DNA barcode to distinguish closely related species. In this study, the complete chloroplast genome of *A. chinense* was obtained using next generation DNA sequencing and bioinformatic analysis, and compared with that from other *Allium* species. The genome is a circular molecule of 152 525 bp with a typical quadripartite structure. Genome annotation identified a total of 116 genes, including 81 protein-coding genes, 31 tRNA genes, and 4 rRNA genes. Analyses of sequences from six *Allium* species showed that the most diverse regions are found in the protein coding regions such as *ndhA* and *ycf1* genes, and in the intergenic regions, such as *ps16-trnQ*, *trnT-trnF*, *ndhF-rpl32*, *rpl32-trnL* and *rpl16-rps3*. A phylogenetic tree was constructed using 58 protein coding sequences from 53 species. All branches showed strong support with bootstrap scores reaching 66%−100%, except those for the Lilium and Paris. Our results suggest that the completed chloroplast genome could solve the classification problems of these species. Using EcoPrimer software, we identified seven markers from the

收稿日期: 2018-09-29; 修回日期: 2018-10-31.

基金项目: 国家自然科学基金资助项目 (81274013, 81473315); 中国医学科学院医学与健康科技创新工程协同创新团队: “本草基因组” (2016-I2M-3-016).

*通讯作者 Tel: 86-10-62811448, E-mail: lfhuang@implad.ac.cn; cliu6688@yahoo.com

DOI: 10.16438/j.0513-4870.2018-0887

chloroplast genomes, which can be used to differentiate congeneric species. In summary, we have sequenced the complete chloroplast genome of *A. chinense*, carried out phylogenetic analysis and identified a series of genus specific DNA barcode sequences. The results have laid the foundation for the systematical determination of the phylogenetic relationship of *Allium* species and the differentiation of species using the genus specific primers.

Key words: *Allium chinense*; complete chloroplast genome; molecular marker; phylogenetic analysis; DNA barcode

藠头 (*Allium chinense*), 又名薤、野葱, 为百合科葱属鳞球茎多年生植物, 起源地和主产区均在亚洲[1]。藠头, 始载于《本草图经》, “味辛、苦, 性温, 具理气宽胸、通阳散结之功”, 与同科植物小根蒜 (*A. macrostemon*) 同为中药薤白的来源, 收录于《中国药典》2015年版, 用于冠心病、心绞痛、胃神经官能症、肠胃炎、久痢等疾病的治疗, 主要成分为甾体皂苷、含硫化合物等。同时, 藠头富含碳水化合物、氨基酸和维生素等有益成分, 具有极高的营养价值, 也被称为“菜中灵芝”[2-4]。1987年国家卫生部将藠头列为药食兼用食品, 被广泛地作为普通食品、功能食品和中药的原料使用[5]。

叶绿体是植物细胞中质体的一种, 普遍存在于陆地植物、藻类和部分原生生物中[6]。叶绿体基因组结构高度保守, 通常包括一个长单拷贝区 (long single copy, LSC) 和一个短拷贝区 (short sequence copy, SSC), 被两个反向重复区 (inverted repeat, IRa、IRb) 划分为经典的四段式结构。大多数被子植物的叶绿体基因组大小在120～160 kb之间[7]。叶绿体基因组序列可以分为蛋白编码区序列和非编码区序列。编码区序列进化速率较慢, 适用于科、目等较高的进化阶元水平的系统进化分析[8]。非编码区序列进化较快, 存在更多的变异信息, 更加适合于科、属等较低分类阶元水平的系统进化分析。非编码区又可以细分为内含子和基因间区, 可应用于种下单元的分子鉴定[9]。

与核基因组相比, 叶绿体基因组具有结构简单、分子量小和拷贝数多等特点, 被广泛用于开发DNA条形码分子标记及系统进化研究。1986年, 烟草 (*Nicotiana tabacum*) 和地钱 (*Marchantia polymorpha*) 叶绿体全基因组被首次报道[10,11]。随着高通量测序和生物信息学分析技术的飞速发展, 获得完整的叶绿体全基因组序列已经成为常规的实验流程, 截止到2018年7月, GenBank已注册超过1 800种植物的叶绿体基因组序列, 进一步拓展了叶绿体基因组在植物分子标记开发和系统发育分析方面的研究与应用。

目前为止, 关于葱属植物基因组水平的系统发育分析未见报道, 部分葱属植物的系统发育位置还不明确[12]。Abugalieva等[13]对葱属植物进行了DNA条形码研究, 发现无法完全区分该属物种。因此亟待开展相关研究解决葱属植物系统发育关系不明确、缺乏有效的分子标记用于植物鉴定的问题[14]。应用高通量测序技术对叶绿体全基因组进行测序为精准鉴定药用植物种质资源、确定其系统发育关系提供了新技术手段。本研究对藠头叶绿体全基因组进行了测序、组装, 并与其他葱属植物叶绿体基因组进行了比较分析, 研究结果为明确葱属植物系统发育关系和开发高分辨率的分子标记用于葱属物种鉴定奠定了基础。

材料与方法

植物材料、DNA提取与测序 藠头新鲜叶片采自中国医学科学院药用植物研究所药用植物园。液氮速冻后放于−80 ℃冰箱待用。用改良CTAB法提取叶片DNA, 琼脂糖凝胶电泳和Qubit3.0 (Thermo Fisher Scientific, 美国) 检测DNA质量和浓度。取合格500 ng DNA构建插入片段大小为500 bp的测序文库, 然后在Illumina Hiseq2000平台上采用双末端测序策略进行测序, 共得到45 789 878 bp的原始序列用于后续分析。

基因组组装与注释 首先于2018年5月, 从GenBank下载1 688条叶绿体基因组序列作为参考序列。利用BLSATN将测序获得的原始读段与参考序列进行相似性分析, E-value=1-e5, 筛选获得相似性片段, 运用SPAdes (SPAdes-3.11.1) 软件组装获得contigs。利用python脚本对contigs片段进行延伸, DNAStar的SeqMan对延伸后的片段进一步组装成完整的叶绿体基因组序列[15,16]。使用Bowtie 2 (v2.0.1) 软件将原始测序读段映射 (mapping) 到组装的基因组序列上, 通过检测序列覆盖度和在contig连接处的读段覆盖情况来评估组装的正确性[17]。通过CPGAVAS软件进行叶绿体基因组注释[18], 并使用Apollo软件对注释信息进行编辑[19]。使用OGDRAW软件绘制叶绿体基因组环状结构示意图[20]。将藠头叶绿体基因组提交至GenBank, 登录号为MK096442。

重复序列分析和基因组比较分析 利用REPuter网站 (https://bibiserv. cebitec. uni-bielefeld. de / reputer/) 预测重复序列[21]。软件参数设置包括: “最小的重复序

列长度=30 bp” 以及“重复序列间的相似度>90%”[22]。利用MISA软件预测叶绿体基因组中的简单重复序列(SSR), 参数设置为: ① 单核苷酸重复次数>8, 二核苷酸和三核苷酸重复次数>4, 四核苷酸、五核苷酸和六核苷酸重复次数>3; ② 2个SSR之间的最小距离设置为100 bp, 若距离小于100 bp, 则两个SSR被当做一个复合微卫星[23]。

以藠头叶绿体基因组为参考, 利用mVISTA软件的Shuffle-LAGAN模式对6种葱属植物的叶绿体基因组进行了序列相似性比对分析[24]。使用DnaSP 5.1进行滑动窗口分析来计算叶绿体基因组之间的核苷酸多样性指数 (Pi), 窗口长度设为600 bp, 步长设为200 bp。

系统发育分析 从GenBank下载来自于50个百合目物种和两个外类群物种的完整叶绿体基因组序列, 分析获得58个共有的蛋白质编码基因: *atpA*、*atpB*、*atpE*、*atpF*、*atpH*、*atpI*、*ccsA*、*clpP*、*matK*、*ndhB*、*ndhC*、*ndhE*、*ndhG*、*ndhH*、*ndhI*、*ndhJ*、*ndhK*、*petA*、*petB*、*petD*、*petG*、*petL*、*petN*、*psaA*、*psaB*、*psaC*、*psaI*、*psbA*、*psbC*、*psbF*、*psbH*、*psbI*、*psbJ*、*psbK*、*psbL*、*psbM*、*psbN*、*psbT*、*rbcL*、*rpl14*、*rpl16*、*rpl2*、*rpl20*、*rpl22*、*rpl23*、*rpoC1*、*rpoC2*、*rps11*、*rps12*、*rps14*、*rps18*、*rps3*、*rps4*、*rps7*、*rps8*、*ycf2*、*ycf3* 和 *ycf4*。使用ClustalW将这58条共有蛋白序列进行全局对比[25]。以波状烟草 (*Nicotiana undulata*) 和东当归 (*Angelica acutiloba*) 作为外类群, 利用PAxML (Rndomized Axelerated Maximum Likelihood) 的最大似然法构建系统发育树[26]。具体的参数设置为“raxmlHPC-PTHREADS - SSE3 - f - n - 100 - m PROTGAMMACPREV/GTRGAMMA -x 551314260 -p 551314260 -o *N_undulata, A_acutiloba* -T20”。进化树分枝的可信度用BOOTSTRAP值即自展值进行评价。自展值的初始值设为1 000次。

EcoPrimer引物设计 从GenBank下载了洋葱 *A. cepa* (NC_024813)、高葶韭 *A. obliquum* (NC_037199)、太白韭 *A. prattii* (NC_037432)、大蒜 *A. sativum* (NC_031829) 和茖葱 *A. victorialis* (NC_037240) 的全基因组序列, 与藠头基因组一起进行分析。首先运行以下命令构建数据库: “ecoPCRFormat.py -g -n Allium.Fo -t Taxonomy Allium.gb”。随后针对构建的数据库运行命令“ecoPrimer -d Allium.Fo -l 100 -L 1000 -e 0 -t species > Allium.Po”发现能够有效区分数据库中的序列的特异性引物。

结果

1 基因组结构特征与注释

藠头叶绿体基因组为典型环状DNA分子, 总长度152 525 bp。具有保守的四分状结构, 包括一个LSC区、一个SSC区和一对IR区, 其长度分别为81 323 bp、18 206 bp和26 498 bp (表1)。藠头叶绿体基因组的整体GC含量为36.68%。其IR区的GC含量 (37.41%) 低于SSC区 (42.86%) 的GC含量, 但高于LSC区 (35.68%) 的GC含量, 表明LSC、SSC和IR区可能来源不同或具有不同选择压力 (图1)。

Table 1 Summary of three *Allium* chloroplast genome features

Latin name	*A. chinensis*	*A. cepa*	*A. sativum*
Total cpDNA size/bp	152 525	153 538	153 172
Length of LSC region/bp	81 323	82 694	82 035
Length of IR region/bp	26 498	26 461	26 563
Length of SSC region/bp	18 206	17 922	18 011
Total GC content/%	36.68	36.81	36.68
GC content in LSC/%	35.68	34.59	34.51
GC content in IR/%	37.41	42.68	42.58
GC content in SSC/%	42.86	29.71	29.15

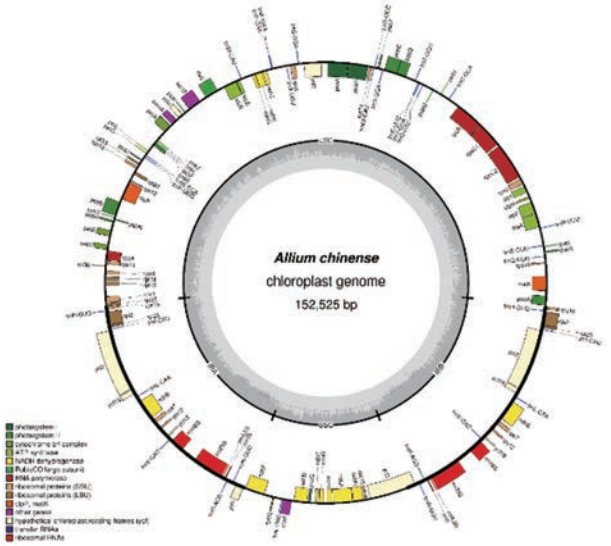

Figure 1 Chloroplast genome map of *Allium chinense*. Genes outside of the map transcripted in the clockwise direction and genes inside of the map transcripted in the counter-clockwise direction. Each color represents the same class of genes

应用CPGAVAS软件对藠头叶绿体基因组进行注释预测了116个基因, 包括蛋白编码基因81个、转运RNA (transfer RNA) 基因31个和核糖体RNA (ribosome RNA) 基因4个 (表2)。其中有9个基因 (*atpF*、*clpP*、*ndhA*、*ndhB*、*rpl2*、*rpoC1*、*ycf1*、*ycf15*、*ycf3*) 含有一个内含子(intron), 另外3个基因 (*clpP*、*ycf1*、*ycf3*) 含有两个内含子。藠头叶绿体全基因组中蛋白编码区 (coding sequence,CDS) 的长度为78 919 bp, 占整个基因组长度的51.7%。rRNA基因的长度为10 296 bp, 占总长度的6.8%。而tRNA基因的长度为2 339 bp, 占整个基因组长度的1.5%。藠头叶绿体基因组非编码区主要包

括内含子、假基因和基因间隔区, 其长度占整个基因组长度的40%。其中由于*ycf1*位于SSC/IRb的边界处, 所以在SSC/IRa区产生了一个假基因。

2 重复序列分析

蔊头重复序列的结构和分布结果见图2。在蔊头叶绿体基因组中共发现50个重复序列满足长度超过30 bp且序列重复的相似性大于90%等两个条件。这50个序列中包括回文重复序列27条、正向重复序列18条、互补重复序列2条、反向重复序列3条。与洋葱、大蒜中的重复序列相比, 蔊头的回文重复、互补重复和反向重复序列的数目均高于其余二者。这些重复序列大部分分布在基因间隔区 (intergenic spacers, IGS), 而且大多数重复序列的长度在30～40 bp之间。蔊头叶绿体基因组中重复序列的类型以A/T为主, 其次为AT/AT, 二者合计占所有重复序列总数的90%以上 (表3)。其中, 大部分是单核苷酸重复, 共检测到171条。大多数的单核苷酸重复类型为A/T单碱基重复。而二核苷酸重复中由AT/AT组成的重复序列最为常见。以上结果和之前报道的结果相符, 即来自叶绿体基因组的SSR通常由短的polyA或polyT重复序列组成, 而由串联的G或C组成的重复序列较为少见[27]。

3 非同义 (K_A) 与同义 (Ks) 替代率分析

为了发现哪些葱属基因在进化过程中被正向或负向选择, 本研究对6个葱属植物叶绿体基因组的65蛋白编码基因进行了K_A/K_S分析。结果显示这65个基因单独的K_A/K_S均小于1, 说明纯化选择 (purifying

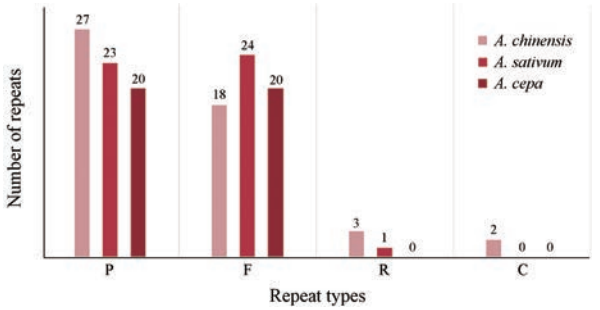

Figure 2 Statistics of repeat sequences detected in the chloroplast genome of *Allium chinense*. P, F, C and R indicate the repeat types P (Palindrome), F (Forward), C (Complement), R (Reverse) respectively

Table 3 Statistics of repeat types detected in the chloroplast genome of *Allium chinense*

Repeat unit	Number of repeated units
A/T	169
C/G	2
AC/GT	1
AG/CT	17
AT/AT	37
AAG/CTT	1
AAT/ATT	1
AAAG/CTTT	1
AAAT/ATTT	5
AATC/ATTG	1
AATG/ATTC	1
ACAG/CTGT	1
AGGG/CCCT	2
ATCC/ATGG	1
AAAAT/ATTTT	1

Table 2 Gene composition in *Allium chinense* chloroplast genome

Category of genes	Group of genes	Name of genes
Self-replication	Large subunit of ribosome	*rpl14, rpl16, rpl2, rpl20, rpl22 rpl23, rpl32, rpl33, rpl36*
	DNA dependent RNA polymerase	*rpoA, rpoB, rpoC1, rpoC2*
	Small subunit of ribosome	*rps11, rps12, rps14, rps15, rps16, rps18, rps19, rps2, rps3, rps4, rps7, rps8*
	rRNA Genes	*rrn4.5, rrn5, rrn16, rrn23*
	tRNA Genes	*trnA-UGC, trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnfm-CAU, trnG-GCC, trnG-UCC, trnH-GUG, trnI-CAU, trnI-GAU, trnK-UUU, trnL-CAA, trnL-UAA, trnL-UAG, trnM-CAU, trnN-GUU, trnP-GGG, trnP-UGG, trnQ-UUG, trnR-ACG, trnR-UCU, trnS-GCU, trnS-GGA, trnS-UGA, trnT-GGU, trnT-UGU, trnV-GAC, trnV-UAC, trnW-CCA, trnY-GUA*
Photosynthesis	Subunits of ATP synthase	*atpA, atpB, atpE, atpF, atpH, atpI*
	Subunits of NADH-dehydrogenase	*ndhA, ndhB, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK*
	Subunits of cytochrome b/f complex	*petA, petB, petD, petG, petL, petN*
	Subunits of photosystem I	*psaA, psaB, psaC, psaI, psaJ*
	Subunits of photosystem II	*psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ*
	Subunit of rubisco	*rbcL*
Other genes	Subunit of acetyl-CoA-carboxylase	*accD*
	c-Type cytochrome synthesis gene	*ccsA*
	Envelop membrane protein	*cemA*
	Protease	*clpP*
	Translational initiation factor	*infA*
	Maturase	*matK*
Unknow	Conserved open reading frames	*ycf*1, *ycf*2, *ycf*3, *ycf*4, *ycf*15, *ycf*68

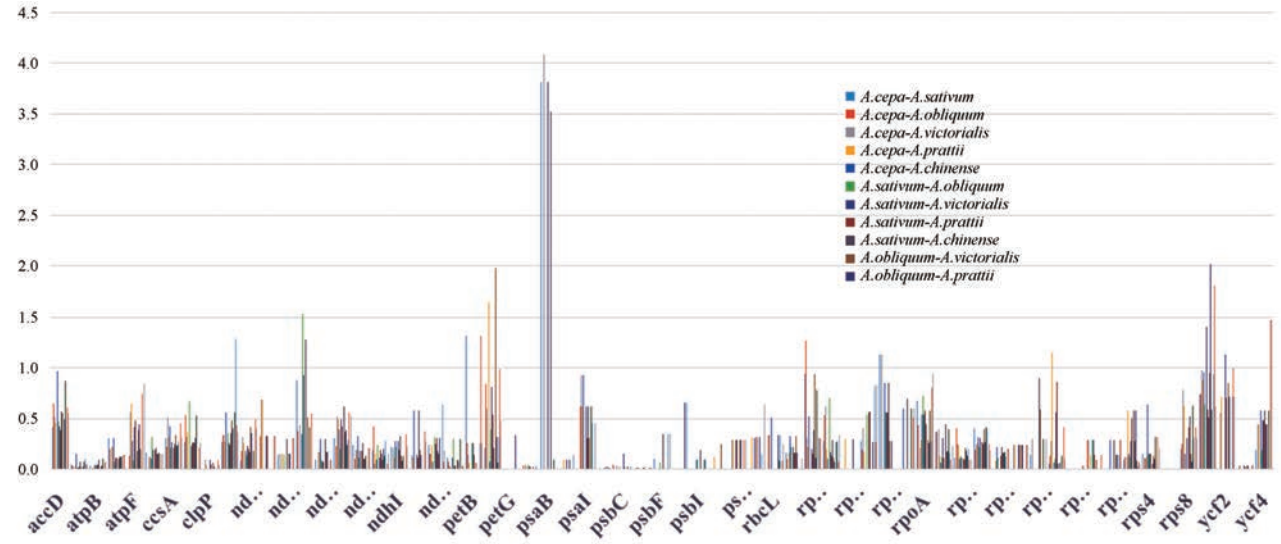

Figure 3 K_A/K_S analysis of 65 common protein coding genes in *Allium* species, *A. chinensis, A. cepa, A. obliquum, A. prattii, A. sativum, A. victorialis*. K_A: Non-synonymous, K_S: Synonymous. 65 common protein coding genes: *accD, atpA, atpB, atpE, atpF, atpI, ccsA, cemA, clpP, matK, ndhA, ndhB, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK, petA, petB, petD, petG, psaA, psaB, psaC, psaI, psbB, psbC, psbD, psbF, psbH, psbI, psbJ, psbK, psbM, rbcL, rpl12, rpl20, rpl22, rpl23, rpl32, rpl33, rpl36, rpoA, rpoB, rpoC1, rpoC2, rps11, rps12, rps14, rps15, rps16, rps18, rps19, rps3, rps4, rps7, rps8, ycf1, ycf2, ycf3* and *ycf4*

selection) 是葱属叶绿体基因组蛋白编码基因碱基突变的主要选择类型。又对每个基因在两两葱属物种中的 K_A/K_S 值进行了分析, 结果见图 3, 发现 *psaB* 基因表现出最大的正选择效应, 说明该基因受到强烈正选择, 且该基因为近期正在快速进化的基因, 对于深化研究物种的进化有着非常重要的意义。

4 比较基因组分析

为了探讨薤头与其他葱属植物叶绿体基因组之间的异同, 将薤头基因组作为参考与其他 5 条葱属植物叶绿体基因组序列进行了相似性比较分析 (图 4)。总的来看, IR 区域比 LSC 和 SSC 区域更加保守, 尤其是蛋白编码区序列高度保守, 而非编码区的序列则有不同程度的区别, 显示葱属植物基因间隔区的进化速率要比基因编码区更快。如图所示, 6 个叶绿体基因组序列整体上高度相似。但是, 薤头与其他 5 个物种在部分基因编码区和基因间区的差异明显, 例如 *ndhA* 和 *ycf1* 基因区, *rps16-trnQ*、*trnT-trnF*、*ndhF-rpl32*、*rpl32-trnL* 和 *rpl16-rps3* 基因间区。从这些区域中有望开发出新的 DNA 条形码用于葱属不同物种的鉴定。

为了进一步发现 6 个葱属植物叶绿体基因组中的高变异位点, 采用滑动窗口分析 (sliding window analysis) 计算了不同区间的核苷酸多样性指数 (Pi, 是核苷酸多样性指数)。如图 5 所示, 6 个葱属植物间的 Pi 平均值为 0.015 347。与序列相似性比较分析一致, IR 区域相较于 LSC 和 SSC 区域更加保守。分析发现 6 个变异热点, 这 6 个热点都位于 LSC 和 SSC 区域, 而且 Pi>0.04。这些位点或许正在经历快速的碱基替换, 其潜在机制有待进一步研究。

5 叶绿体基因组系统发育分析

利用来自于 51 个百合目物种和 2 个外类群物种的 58 个共有蛋白序列构建 ML 系统进化树 (图 6)。结果显示, 系统进化树与已知的物种进化关系一致。例如 *Aloe*、*Polygonatum*、*Hosta*、*Allium*、*Paris*、*Fritillaria*、*Dioscorea* 等属的物种都聚在了一起。薤头与其他葱属植物聚为一支, 与洋葱 (*A. cepa*)、大蒜 (*A. sativum*) 的亲缘关系比与茖葱 (*A. victorialis*)、太白韭 (*A. prattii*) 和高葶韭 (*A. obliquum*) 近。进化树中各个节点的 bootstrap 分值较高, 表明应用叶绿体全基因组序列可以有效解决百合目下各科、属、种的分类问题。然而, 同属部分物种对应 bootstrap 分值较小, 表明该分枝的可信度较低, 如重楼属、百合属对应的分枝。这些物种的分类仍然存在着问题, 需要后期进一步研究。

6 EcoPrimer 引物设计结果

利用 EcoPrimer 对葱属 6 个植物叶绿体基因组序列的潜在通用条码区进行引物设计, 结果见表 4。

讨论

葱属 *Allium* 为百合科第一大属, 全世界共有约 970 种, 我国有 9 组 99 种, 目前, 葱属植物遗传背景工作研究有限, 除本文报道的薤头外, 只有 5 个葱属物种的叶

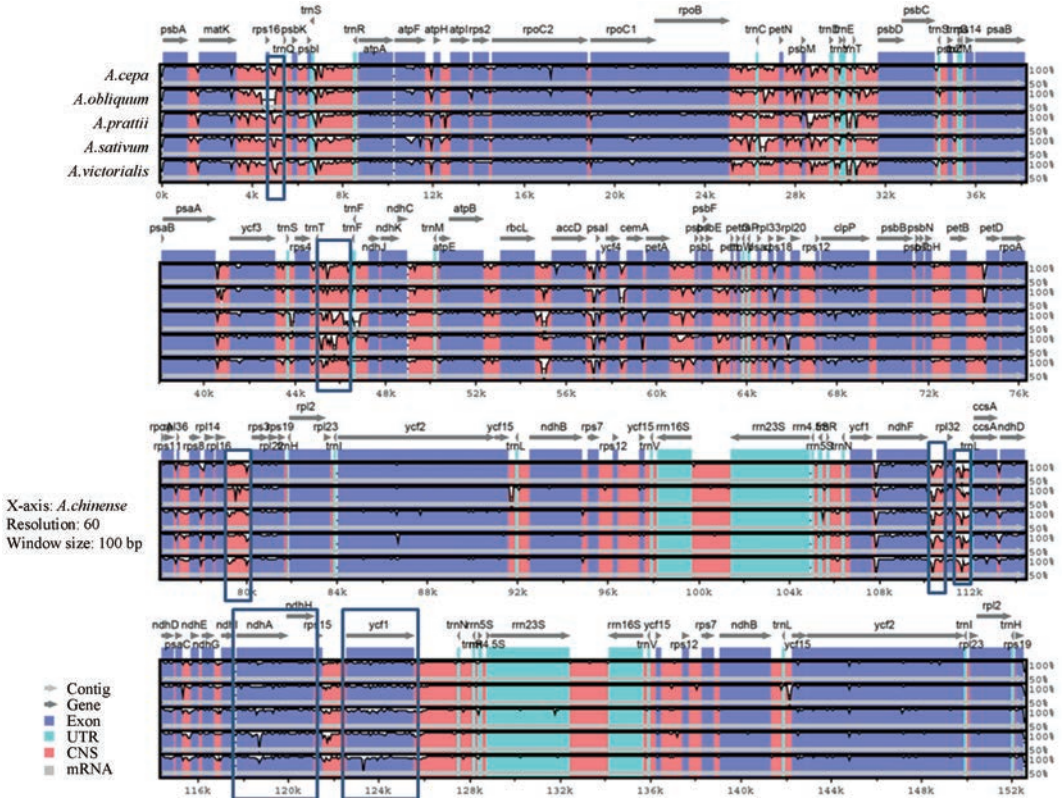

Figure 4 Identify plot comparing the chloroplast genomes of six species using *A. chinense* as the reference sequence. The vertical scale, ranging from 50% to 100%, indicates the percentage of identity calculated in sliding windows. The horizontal axis indicates the coordinates within the chloroplast genome. Different colors correspond to the types of the genome regions. Blue: Regions coding for proteins; Pink: Regions that are non-coding; Light blue: Regions coding for tRNAs and rRNAs

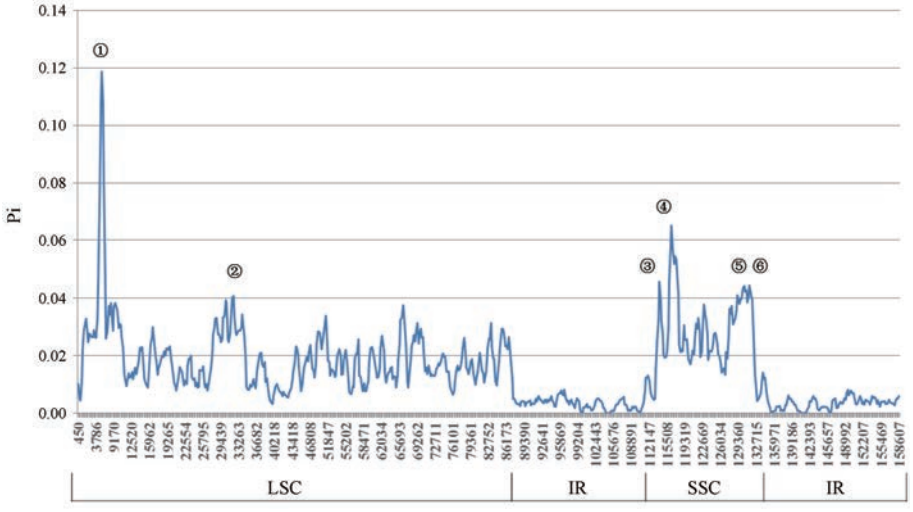

Figure 5 Sliding window analysis of the whole chloroplast genomes of six *Allium* species. Window length: 800 bp; step size: 200 bp. X-axis: Position of the midpoint of a window; Y-axis: Value of (Nucleotide diversity) Pi of each windows

绿体基因组序列得以报道[28]。与该属所具有的物种多样性水平相比, 基础研究工作明显不足。叶绿体基因组含有丰富分子系统发育信息, 在阐明该科、属内物种间的系统进化关系、发现DNA条形码候选分子标记等方面具有重要意义。

藠头叶绿体基因组大小为总长度152 525 bp, 与已发表葱属植物叶绿体基因组大小相似。该基因组呈典型环状DNA分子, 具有保守的四分状结构[29,30]。对6个葱属植物的叶绿体基因组进行比较分析, 发现*ndhA*和*ycf1*基因的变异速率快于其他基因, 提示这两

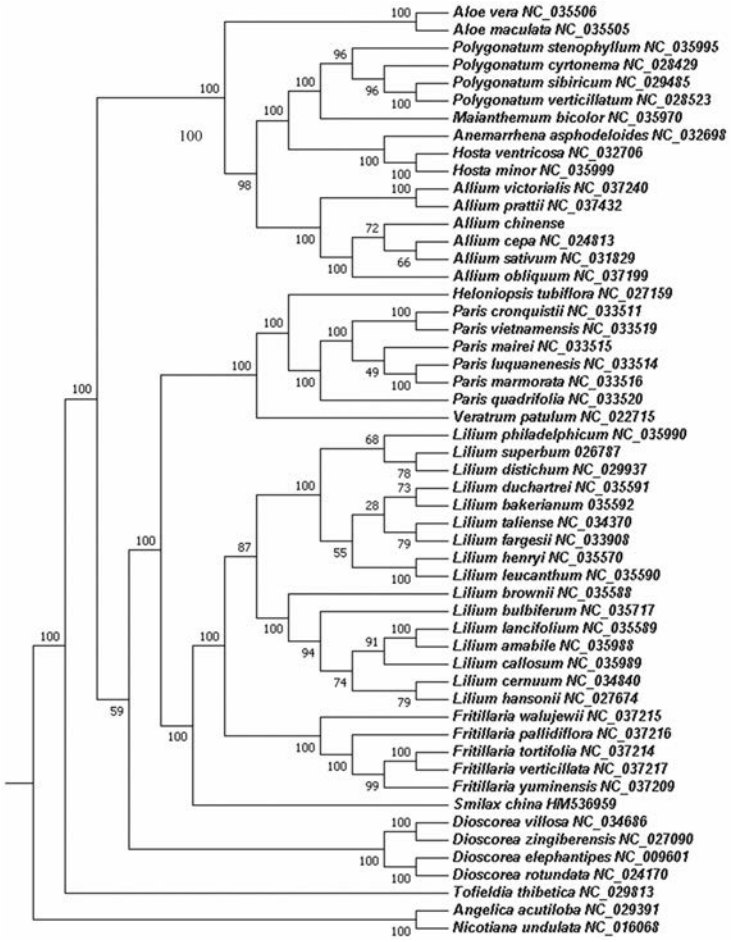

Figure 6 Phylogenetic tree constructed based on 58 protein-coding genes from species. *atpA, atpB, atpE, atpF, atpH, atpI, ccsA, clpP, matK, ndhB, ndhC, ndhE, ndhG, ndhH, ndhI, ndhJ, ndhK, petA, petB, petD, petG, petL, petN, psaA, psaB, psaC, psaI, psbA, psbC, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, rbcL, rpl14, rpl16, rpl2, rpl20, rpl22, rpl23, rpoC1, rpoC2, rps11, rps12, rps14, rps18, rps3, rps4, rps7, rps8, ycf2,* y*cf3* and *ycf4*

Table 4 Primer designed by EcoPrimer

Name	Forward primer sequence	Reverse primer sequence
ndhA	TTCTATTCTTTATAGGTA	TGAGTCACAGTCAAAATA
ycf1	GCTTCTTTCTTGCTCTTA	AATTATCTAGAAAAAAGA
rps16-trnQ	TTGTTTTCATGAGTATGA	AATCATTTTTTTCTATAT
trnT-trnF	AGGCCCTTTTAACTCAGT	GCTGCGGGTTCGAGCCCC
ndhF-rpl32	AACACCAATCCCATTTAT	TCAAAAAATACATATCAA
rpl32-trnL	AGTTCCAAAAAAACGTAC	TCTAGAATCCGATATAGT
rpl16-rps3	CTGAAATAACAAATTGAG	CCCGGTCTACGAATTTAT

个基因可用于物种鉴定及亲缘关系的梳理。进一步发现基因间区 *rps16-trnQ*、*trnT-trnF*、*ndhF-rpl32*、*rpl32-trnL* 和 *rpl16-rps3* 变异幅度大, 有望发现葱属特异的DNA分子标记用于物种鉴定。利用EcoPrimer软件发现了7个潜在标记, 为进一步开发葱属特异性分子标记奠定了基础。简单重复序列 (SSR) 广泛分布于各种真核生物的基因组中, 在植物基因组中含量丰富[31]。因其重复单位的数目存在高度变异, 常常造成相关序列的多态性, 已广泛用于种群遗传学、系统发育学研究中[32]。SSR标记与传统标记方法相比, 具有数量丰富、覆盖度高、具多等位基因、种属特异性强等优点, 然而其开发成本较高, 因此未被广泛应用。本研究从菖头叶绿体基因组中发现了241个SSR, 包括171个单核苷酸、55个二核苷酸、2个三核苷酸、12个四核苷酸和1个五核苷酸。为未来筛选SSR标记奠定基础。

DNA碱基突变根据对其编码氨基酸的影响可以分为两类: 同义突变 (K_S) 和非同义突变 (K_A), 同义突变不导致氨基酸改变, 其频率用 K_S 表示, 非同义突变

导致氨基酸改变, 其频率用 K_A 表示。非同义 (K_A) 和同义 (K_S) 替代率 (K_A/K_S) 是揭示进化率和自然选择压力的重要指标[33]。一般来讲, K_A/K_S >1 时, 则认为有正选择效应, 而 K_A/K_S <1 时, 则认为有纯化选择作用。对大多数蛋白来说, 同义核苷酸替换出现的频率比非同义核苷酸替换出现的频率高, 所以对大多数蛋白编码区来说 K_A/K_S 值通常小于 1[34]。通过对葱属植物进行非同义 (K_A) 和同义 (K_S) 替代率 (K_A/K_S) 分析, 发现 psaB 基因表现出最大的正选择效应, 说明该基因受到强烈正选择, 且预示该基因为近期正在快速进化的基因, 对下一步研究葱属物种的进化有着重要的意义。

叶绿体基因组序列含有丰富的遗传信息, 以波状烟草和东当归为外类群, 构建基于叶绿体全基因组的系统发育树, 表明叶绿体基因组可以有效解决百合目下各科、属、种的分类问题, 在已有叶绿体基因组序列的葱属植物内可成功区分。但因受葱属物种叶绿体全基因组序列信息不全的影响, 对葱属鉴别和系统发育研究仍需后期进一步研究、验证。

基于本研究结果, 下一步的工作可能涉及以下几个方向: 从位于基因间区、变异幅度最大的区间 *ndhA*、*ycf1*、*rps16-trnQ*、*trnT-trnF*、*ndhF-rpl32*、*rpl32-trnL* 和 *rpl16-rps3* 中发现新的分子标记, 评价其鉴别效率; 这些分子标记还可以与来自于核基因组的分子标记序列结合起来一起分析, 进一步提高对葱属物种的鉴别效率; 对叶绿体基因组中发现的 SSR 标记进行进一步分析, 为藠头的种质资源开发、物种多样性研究及分子育种研究提供技术手段; 对表现出强烈正选择的 *psaB* 基因在葱属植物的其他物种中进行深入研究, 确认该基因在属内是否显示强烈的正选择, 通过该基因了解葱属物种的进化过程。

References

[1] Mann LK, Stearn WT. Rakkyo or Ch'iao T'ou (*Allium Chinense* G. Don, Syn. A. bakeri Regel) a little known vegetable Crop [J]. Econ Bot, 1960, 14: 69-83.

[2] Lim TK. Allium chinense [M]// Edible Medicinal and Non Medicinal Plants. Dordrecht: Springer, 2015: 204-209. https://doi.org/10.1007/978-94-017-9511-1_5.

[3] Peng JP, Yao XS, Tezuka Y, et al. Furostanol glycosides from bulbs of *Allium chinense* [J]. Phytochemistry, 1996, 41: 283-285.

[4] Baba M, Ohmura M, Kishi N, et al. Saponins isolated from Allium chinense G. Don and antitumor-promoting activities of isoliquiritigenin and laxogenin from the same drug [J]. Biol Pharm Bull, 2000, 23: 660-662.

[5] Zou ZM, Yu DQ, Cong PZ. Research progress in the chemical constituents and pharmacological actions of *Allium* species [J]. Acta Pharm Sin (药学学报), 1999, 34: 395-400.

[6] Meyer CP, Paulay G. DNA barcoding: error rates based on comprehensive sampling [J]. PLoS Biol, 2005, 3: e422.

[7] 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.

[8] Li XW, Hu ZG, Lin XH, et al. High-throughput pyrosequencing of the complete chloroplast genome of *Magnolia officinalis* and its application in species identification [J]. Acta Pharm Sin (药学学报), 2012, 47: 124-130.

[9] Shaw J, Lickey EB, Schilling EE, et al. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III [J]. Am J Bot, 2007, 94: 275-288.

[10] Sugiura M, Shinozaki K, Zaita N, et al. Clone bank of the tobacco (*Nicotiana tabacum*) chloroplast genome as a set of overlapping restriction endonuclease fragments: mapping of eleven ribosomal protein genes [J]. Plant Sci, 1986, 44: 211-217.

[11] Ohyama K, Fukuzawa H, Kohchi T, et al. Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha*, chloroplast DNA [J]. Nature, 1986, 322: 572-574.

[12] Cowan RS, Chase MW, Kress WJ, et al. 300 000 species to identify: problems, progress, and prospects in DNA barcoding of land plants [J]. Taxon, 2006, 55: 611-616.

[13] Abugalieva S, Volkova L, Genievskaya Y, et al. Taxonomic assessment of Alliumspecies from Kazakhstan based on ITS and matK markers [J]. BMC Plant Biol, 2017, 17 (Suppl 2): 258.

[14] Tang YC. Notes on changes in classification of *Liliaceae* (S. L.) and perspective in China [J]. Acta Phytotaxon Sin (植物分类学报), 1995, 33: 1-26.

[15] Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing [J]. J Comput Biol, 2012, 19: 455-477.

[16] Burland TG. DNASTAR's Lasergene sequence analysis software [J]. Methods Mol Biol, 2000, 132:71-91.

[17] Langmead B. Aligning short sequencing reads with Bowtie [J]. Curr Protoc Bioinformatics, 2010, Chapter 11: Unit 11.7. DOI: 10.1002/0471250953.bi1107s32.

[18] 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 [J]. BMC Genomics, 2012, 13: 715-715.

[19] Lewis SE, Searle S, Harris N, et al. Apollo: a sequence annotation editor [J]. Genome Biol, 2002, 3: research0082.1-82.14.

[20] Lohse M, Drechsel O, Bock R. Organellar Genome DRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes [J]. Curr Genetics, 2007, 52: 267-274.

[21] 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.

[22] Martin GE, Rousseaugueutin M, Cordonnier S, et al. The first complete chloroplast genome of the Genistoid legume *Lupinus luteus*: evidence for a novel major lineage-specific rearrangement and new insights regarding plastome evolution in the legume family [J]. Ann Bot, 2014, 113: 1197.

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

[24] Frazer KA, Pachter L, Poliakov A, et al. VISTA: computational tools for comparative genomics [J]. Nucleic Acids Res, 2004, 32 (Web Server issue): W273.

[25] Corpet F. Multiple sequence alignment with hierarchical clustering [J]. Nucleic Acids Res, 1988, 16: 10881-10890.

[26] Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies [J]. Bioinformatics, 2014, 30: 1312-1313.

[27] Wang S, Shi C, Gao LZ. Plastid genome sequence of a wild woody oil species, *Prinsepia utilis*, provides insights into evolutionary and mutational patterns of rosaceae chloroplast genomes [J]. PLoS One, 2013, 8: e73946.

[28] He XJ, Ge S, Xu JM, et al. Phylogeny of Chinese *Allium* using PCR-RFLP analysis [J]. Sci China C (中国科学: 生命科学), 2000, 30: 183-191.

[29] Kim S, Park JY, Yang T. Comparative analysis of the complete chloroplast genome sequences of a normal male-fertile cytoplasm and two different cytoplasms conferring cytoplasmic male sterility in onion (*Allium cepa* L.) [J]. J Hort Sci Biotechnol, 2015, 90: 459-468.

[30] Jin FY, Xie DF, Zhou SD, et al. Characterization of the complete chloroplast genome of *Allium prattii* [J]. Mitochondrial DNA Part B, 2018, 3: 153-154.

[31] Mccouch SR, Teytelman L, Xu Y, et al. Development and mapping of 2240 new SSR markers for rice (*Oryza sativa* L.) (supplement) [J]. DNA Res, 2002, 9: 257-279.

[32] Zietkiewicz E, Rafalski A, Labuda D. Genome fingerprinting by simple sequence repeat (SSR)-anchored polymerase chain reaction amplification [J]. Genomics, 1994, 20: 176-183.

[33] Yang Z, Nielsen R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models [J]. Mol Biol Evol, 2000, 17: 32-43.

[34] Makalowski W, Boguski MS. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2 820 orthologous rodent and human sequences [J]. Proc Natl Acad Sci U S A, 1998, 95: 9407-9412.