秦岭柴胡叶绿体全基因组特征解析及柴胡属种间系统发育关系分析

陈锦豪[1#], 赵 晨[1#], 高 静[1,2], 李依民[1,2], 张 岗[1,2], 陈 莹[1,2], 颜永刚[1,2], 张明英[1,2*]

(1. 陕西中医药大学药学院, 陕西省秦岭中草药应用开发工程技术研究中心, 陕西 西安 712046; 2. 陕西中医药大学, 陕西省中医药管理局“秦药”研发重点实验室, 陕西 西安 712046)

摘要: 秦岭柴胡 *Bupleurum giraldii* (H. Wolff) Koso-Pol. 为伞形科 Apiaceae 柴胡属 *Bupleurum* L. 多年生草本植物, 全草可药用, 是陕西省重点保护高山珍稀药用植物。本研究基于基因组浅层测序 (genome skimming, GS) 方法测序获得秦岭柴胡叶绿体全基因组序列, 对其进行了组装、注释和特征解析, 并与同属已发表物种叶绿体全基因组进行了比较和种间系统发育关系分析。秦岭柴胡叶绿体全基因组呈保守的环状四分体结构, 长度为 155 761 bp, 由一个大单拷贝区 (large single copy, LSC; 85 663 bp), 一个小单拷贝区 (small single copy, SSC; 17 492 bp) 和一对反向重复区 (inverted repeats, IRa/IRb; 26 303 bp) 构成; 共注释得到 129 个基因, 包括 84 个蛋白编码基因、37 个 tRNA 基因和 8 个 rRNA 基因, 其中 6 个蛋白编码基因、7 个 tRNA 基因和 4 个 rRNA 基因在 IRs 区发生重复。从秦岭柴胡叶绿体全基因组序列中共检测出 66 个简单重复序列 (simple sequence repeats, SSRs), 34 个串联重复序列 (tandem repeats) 和 31 个散在重复序列 (dispersed repeats)。密码子使用偏好性 (codon usage bias, CUB) 分析结果显示, 当相对同义密码子使用度 (relative synonymous codon usage, RSCU) 大于 1 时, 偏向以 A 或 T 碱基结尾, 而小于 1 时, 则偏向以 C 或 G 碱基结尾; 中性绘图、ENC-plot 和 PR2-plot 分析结果均表明, 秦岭柴胡叶绿体基因组密码子使用偏好性主要受自然选择影响。柴胡属植物种间叶绿体全基因组序列长度、GC 含量、基因组成及排列顺序高度保守, IRs 区未检测到明显的扩张或收缩; 序列的种间变异主要发生在非编码的基因间区, 且 LSC 区、SSC 区序列的变异程度显著高于 IRs 区, 共筛选到 11 条种间高变异序列, 分别位于 LSC 和 SSC 区。系统发育分析结果显示, 柴胡属所有物种以 BS_{ML} = 100%, PP = 1 的支持率聚类为一个单系, 属下划分为柴胡亚属 (subgenus *Bupleurum*) 和 subgenus *Penninervia* 2 个分支, 秦岭柴胡位于柴胡亚属, 与 *B. falcatum* L. 系统发育关系最近缘。本研究首次对秦岭柴胡叶绿体全基因组特征进行了全面解析, 澄清了秦岭柴胡在柴胡属内的系统发育位置, 研究结果将为柴胡属药用植物种质资源保护和合理开发利用提供理论基础。

关键词: 柴胡属; 秦岭柴胡; 叶绿体基因组; 序列变异; 系统发育

中图分类号: R931 **文献标识码**: A **文章编号**: 0513-4870(2026)07-2757-16

Characterization of the complete plastid genome of *Bupleurum giraldii* (H. Wolff) Koso-Pol. and phylogenetic relationships of *Bupleurum* L.

CHEN Jin-hao[1#], ZHAO Chen[1#], GAO Jing[1,2], LI Yi-min[1,2], ZHANG Gang[1,2], CHEN Ying[1,2], YAN Yong-gang[1,2], ZHANG Ming-ying[1,2*]

(*1. School of Pharmacy, Shaanxi University of Traditional Chinese Medicine, Shaanxi Qinling Engineering Research Center for Application and Development of Chinese Herbal Medicine, Xi'an 712046, China; 2. Shaanxi University of Traditional Chinese Medicine, Shaanxi Provincial Administration of Traditional Chinese Medicine Key Laboratory of "Qin Medicine", Xi'an 712046, China*)

Abstract: *Bupleurum giraldii* (H. Wolff) Koso-Pol. is a perennial herbaceous plant belonging to the *Bupleurum*

收稿日期: 2026-01-22; 修回日期: 2026-04-06.

基金项目: 陕西省自然科学基础研究计划项目 (2019JQ-876, 2022JM-458); 陕西省教育厅青年创新团队科学研究计划项目 (25JP045); 秦创原中医药产业创新聚集区项目 (L2024-QCY-ZYYJJQ-X77, L2024-QCY-ZYYJJQ-X48); 咸阳市重点研发计划 (L2023-ZDYF-SF-017); 陕西中医药大学“秦药”品质评价及资源开发学科创新团队项目 (2019-QN01).

#共同第一作者

*通讯作者 Tel: 86-29-38185165, E-mail: zhangmy@sntcm.edu.cn

DOI: 10.16438/j.0513-4870.2026-0103

L. (Apiaceae). It is recognized as a provincially protected rare medicinal plant in Shaanxi with the whole plant has acknowledged medicinal value. In this study, the complete plastid genome sequence of *B. giraldii* was sequenced and characterized, and comparative and phylogenetic analyses were then conducted alongside other published *Bupleurum* plastid genome sequences. The whole plastid genome of *B. giraldii* was 155 761 bp in length, which exhibited the conserved circular quadripartite structure consisting of a large single copy (LSC, 85 663 bp), a small single copy (SSC, 17 492 bp), and a pair of inverted repeats (IRa/IRb, 26 303 bp). A total of 129 genes were annotated, including 84 protein-coding genes (PCGs), 37 transfer RNA (tRNA) genes and 8 ribosomal RNA (rRNA) genes, of which, 6 protein-coding genes, 7 tRNA genes and 4 rRNA genes were duplicated in the IR regions. A total of 66 simple sequence repeats (SSRs), 34 tandem repeats, and 31 dispersed repeats were detected. In the plastid genome of *B. giraldii*, codons with the relative synonymous codon usage (RSCU) value greater than 1 showed a tendency to end in adenine (A) or thymine (T), whereas those with an RSCU value less than 1 tended to end in cytosine (C) or guanine (G). Collectively, neutral plotting, ENC-plot, and PR2-plot analyses all revealed that natural selection might be the predominant factor shaping this codon usage bias. The plastid genome of *Bupleurum* species exhibited highly conserved gene order and content, along with little variation in overall size and GC content, no significant expansion or contraction of the IR regions was detected. Interspecific sequence variation mainly occurred in the non-coding intergenic regions, and sequence divergence in the LSC and SSC regions was significantly higher than that in the IR regions. Eleven divergence hotspot sequences were identified, located in the LSC and SSC regions respectively. Phylogenetic inferences inferred that all *Bupleurum* species clustered into a monophyletic clade with BS_{ML} = 100%, PP = 1, and which could be further divided into two independent lineages corresponding to subgenus *Bupleurum* and subgenus *Penninervia*. *Bupleurum giraldii* located in subgenus *Bupleurum* and showed the closest relationship to *B. falcatum* L. The complete plastid genome of *B. giraldii* was comprehensively investigated for the first time and its phylogenetic position in *Bupleurum* was elucidated with high resolution in this study, these results will provide basis for the conservation and sustainable utilization of the germplasm resources of medicinal *Bupleurum* species.

Key words: *Bupleurum* L.; *Bupleurum giraldii* (H. Wolff) Koso-Pol.; plastid genome; sequence variation; phylogeny

柴胡属 *Bupleurum* L. 隶属于伞形科 Apiaceae, 目前全球记载柴胡属植物约 224 种 (https://powo.science.kew.org/), 中国分布有 42 种, 其中 22 种为中国特有[1], 36 种具有药用价值[2], 以根或带根全草入药, 含有皂苷、挥发油、有机酸、黄酮、甾醇等多种药用活性成分, 具有抗炎、抗肿瘤、神经调节、免疫调节等功效[3,4], 被广泛用于中医临床和现代药物开发研究。

由于丰富的物种多样性, 复杂的物种形成和演化历史, 柴胡属自建立以来其种间划分和系统发育关系问题长期受到关注[5]。Shu 等[6]基于形态、解剖及花粉性状对中国药用柴胡 14 种、2 变种和 1 个变型的分析结果首次提出将柴胡属划分为大叶柴胡亚属 subgenus *Longifolia* (Wollf) Yuan (仅包含大叶柴胡 *B. longiradiatum* Turcz. 1 个物种) 和真柴胡亚属 subgenus *Eubupleura* (Briq.) Yuan, 后者又可分为大苞组 section *Ranunculoidea* (Wolff) Yuan 和小苞组 section *Falcata* (Wolff) Yuan。Neves 等[7]基于核糖体 DNA ITS 序列对柴胡属 32 个物种 (35 个代表个体) 的分子系统发育分析结果将柴胡属划分为柴胡亚属 subgenus *Bupleurum* 和 *Penninervia* 亚属 (subgenus *Penninervia*)。Wang 等[8]结合形态特征和分子证据 (ITS、*trn*H-*psb*A 及 *mat*K 序列) 的系统发育分析结果与 Neves 等[7]的结果一致, 同样将柴胡属划分为柴胡亚属和 subgenus *Penninervia*, 其中, 中国分布的物种均属于柴胡亚属。然而, 由于有限的分子片段所能提供的物种演化信息位点不足, 利用 ITS、*mat*K、*rbc*L、*trn*H-*psb*A 和 *rps*16 等片段单独或联合分析所得的柴胡属种间系统发育关系支持率相对较低[5,7,8], 且不同分析结果所得的系统发育关系间存在冲突, 部分近缘物种界定模糊, 这不仅限制了对柴胡属分类修订、物种分化及种间亲缘演化关系的深入理解, 也直接影响了柴胡属药用植物种质资源的精准鉴定和分类保护。

秦岭柴胡 *B. giraldii* (H. Wolff) Koso-Pol. 又称金柴胡, 主要分布于我国陕西、山西、宁夏、青海等地, 生长于海拔 2 800～3 300 米的高山草地, 全草可入药, 具有发表祛风、清肝利胆、清心火、通经之功效[2,9], 是陕西省重点保护的高山珍稀药用植物, 也是陕西秦巴山区常用中草药[10]。秦岭柴胡的药用功效与同属北柴胡 *B. chinense* DC.、狭叶柴胡 *B. scorzonerifolium* Willd. (中药材柴胡的基原植物) 存在明显差异, 是柴胡属药用资源的重要补充。野外调查发现, 过度采挖导致生

境破碎, 加之自身繁殖更新困难等因素, 秦岭柴胡野生居群的数量和分布范围持续缩减, 逐渐呈现零散分布的濒危态势, 亟需加强对其野生种质资源开展保护。明确其遗传背景与系统发育位置, 是开展遗传多样性评价、种质资源精准鉴定及制定有效保护策略的重要前提。然而, 秦岭柴胡在柴胡属的系统发育位置较少受到关注。Wang 等[8]基于核基因间区 ITS、叶绿体 *trn*H-*psb*A 和 *mat*K 序列对中国柴胡属的系统发育分析结果显示, 秦岭柴胡位于柴胡亚属, 与贵州柴胡 *B. kweichowense* R. H. Shan (*trn*H-*psb*A+*mat*K) 或贵州柴胡+空心柴胡 *B. longicaule* var. *franchetii* H. Boissieu (ITS+*trn*H-*psb*A+*mat*K) 构成姐妹关系, Song 等[5]基于 ITS、*mat*K、*trn*H-*psb*A 和 *rbc*L 序列对东亚地区分布的柴胡属种间系统发育分析结果亦将秦岭柴胡与贵州柴胡+空心柴胡聚类为一个单系分支, 但支持率均相对较低 (BS_{ML} = 67%～78%, PP = 0.75～0.95)。系统发育位置的不确定阻碍了对秦岭柴胡这一狭域分布的高山珍稀药用植物种质资源的精准评估与保护, 同时也将制约对其药用资源的合理开发利用。

叶绿体不仅是绿色植物及藻类等真核自养生物细胞进行光合作用的重要场所, 同时其编码的许多基因还直接或间接参与萜类、酚类等药用次生代谢产物的合成与调控[11]。作为一种半自主性细胞器, 其遗传物质叶绿体基因组 (chloroplast genome) 可独立于核基因组之外进行自我复制和遗传[12]。被子植物叶绿体基因组结构稳定, 基因组成和排列顺序高度保守, 且基因组较小, 在细胞中常存在多个拷贝, 易于测序分析。同时, 由于其核苷酸替换率相对较低和单亲遗传特性, 使得叶绿体基因组序列成为植物类群种间划分、分子鉴定和系统发育关系重建研究的理想分子标记[13-16]。解析秦岭柴胡叶绿体基因组的序列组成与结构特征, 明确其与同属近缘物种叶绿体基因组间的差异, 可为厘清其系统发育位置、完善柴胡属分类修订提供可靠的分子证据; 同时, 挖掘与次生代谢相关的基因组特征, 将进一步为秦岭柴胡药用特性分子基础的揭示提供线索, 促进其种质资源的保护和药用资源的合理开发利用。目前已有 19 950 个物种共 42 784 条叶绿体基因组序列在国家生物信息中心叶绿体基因组综合数据库 (CGIR, https://ngdc.cncb.ac.cn/cgir/) 收录和开放获取 (截至 2026 年 1 月 5 日)。柴胡属北柴胡[17]、长茎柴胡 *B. longicaule* Wall. ex DC[18]、云南柴胡 *B. yunnanense* Franch.[19]、马尔康柴胡 *B. malconense* R. H. Shan & Yin Li[20]等多个物种的叶绿体基因组特征已被解析, 并基于叶绿体全基因组序列开展了柴胡属部分物种的种间系统发育关系重建, 所得系统发育树的支持率较以往基于 ITS、*mat*K、*rbc*L、*trn*H-*psb*A 和 *rps*16 等序列的分析结果得到显著提升。然而, 秦岭柴胡的叶绿体全基因组尚未见报道。

本研究以秦岭柴胡为研究对象, 基于二代高通量基因组浅层测序 (genome skimming, GS) 方法测序获得叶绿体全基因组序列, 对其进行组装、注释和特征解析, 并与同属已发表物种的叶绿体全基因组进行序列变异比较和系统发育重建分析, 旨在全面解析秦岭柴胡叶绿体全基因组特征, 探讨柴胡属植物叶绿体全基因组序列的种间变异模式, 并澄清秦岭柴胡在柴胡属内的系统发育位置, 为秦岭柴胡的居群划分、种质资源鉴定及珍稀物种保护策略制定提供理论基础。

材料与方法

实验材料 野生秦岭柴胡新鲜幼嫩叶片材料采自陕西省宝鸡市太白县小文公庙附近 (107.803 268 00°E、33.990 082 00°N), 硅胶快速脱水干燥保存, 同时采集凭证标本。凭证标本由陕西中医药大学药学院王继涛高级实验师鉴定 (采集号: Bup_cp21, 采集人: 董栓录), 保存于陕西中医药大学中药标本馆。

基因组 DNA 提取与测序 利用高效植物基因组 DNA 提取试剂盒 (DP350, TIANGEN, 北京) 提取基因组总 DNA, 1% 的琼脂糖凝胶电泳检测 DNA 质量, Qubit 3.0 (Thermo Fisher Scientific, 美国) 荧光定量仪检测 DNA 浓度。检测合格的基因组总 DNA 利用 NEBNext® Ultra™ II DNA 文库制备试剂盒 (New England Biolabs, 美国) 构建长度约为 500 bp 的插入片段文库, 利用 DNBSEQ-T7 测序平台 (深圳华大智造科技股份有限公司) 进行 2×150 bp 双端测序 (paired-end)。测序数据质量控制标准为: Q20 ≥ 95%、Q30 ≥ 90%, N 碱基比例 < 1%, 序列重复率< 10%, 测序深度≥ 100×。测序工作由深圳华大基因股份有限公司武汉分公司完成。

叶绿体全基因组序列组装、拼接与注释 测序得到约 6.90 Gb 共 40 640 242 条原始序列 (raw reads), 平均碱基覆盖度为 4 515.551×。利用 NGS QC Toolkit v.2.3.3[21]过滤去除接头及低质量序列 (参数设置为 cutoff value for percentage of read length = 80, cut-off value for PHRED quality = 30), 最终得到 40 259 544 条 (占比 99.06%) 高质量序列 (clean reads)。根据 Song 等[5]、Wang 等[8]基于 ITS、*trn*H-*psb*A、*mat*K 及 *rbc*L 序列对柴胡属的分子系统发育分析结果, 北柴胡与秦岭柴胡、贵州柴胡及空心柴胡等物种亲缘关系较近, 聚类为一个单系。目前, 北柴胡叶绿体全基因组已有研究报道解析, 并在 GenBank 数据库开放获取。据此, 以 GenBank 数据库中收录的北柴胡叶绿体全基因组序

列 (检索号: MN756857) 为参考, 利用 GetOrganelle v.1.7.5[22]对过滤后的序列进行组装 (参数设置为get_organelle_from_reads.py -1 R1.fq -2 R2.fq - o cp_output -R 15 -k 21,45,65,85,105 -F embplant), 获得序列重叠群 (contigs), 再利用 Bandage v.0.8[23]将组装获得的序列重叠群进行拼接并去除冗余序列, 并在 Geneious v.8.0.2[24]中通过与参考序列比对确定两个反向重复区的方向, 最终获得完整的环状四分体叶绿体全基因组。利用CPGAVAS2[25]对组装拼接完成的秦岭柴胡叶绿体全基因组序列进行注释, 并在 Geneious v.8.0.2 中检查校正注释结果, 利用在线软件 tRNAscan-SE[26] (https://toolkit. engineering. ucsc. edu/website-support/maintenance/) 确定 tRNA 基因的边界。利用 OGDRAW[27] (https://chlorobox. mpimp-golm. mpg. de/OGDraw. html) 在线绘制叶绿体全基因组图谱。组装并注释完成的叶绿体全基因组提交至 GenBank 数据库 (检索号: MW442884)。

叶绿体全基因组重复序列检测 秦岭柴胡叶绿体全基因组中的散在重复序列 (dispersed repeats), 包括正向重复 (forward repeat)、反向重复 (reverse repeat)、互补重复 (complementary repeats) 以及回文重复 (palindromic repeats), 利用 REPuter[28] (https://bibiserv.cebitec.uni-bielefeld.de/reputer) 进行分析检测, 参数设置为: 最小重复序列长度 = 30 bp, 重复序列间的相似度 > 90% (Hamming distance = 3, maximum computed repeats = 90, minimal repeat size = 30)。串联重复序列 (tandem repeats) 利用 Tandem Repeat Finder (TRF)[29] (https://tandem.bu.edu/trf/trf.unix.help.html) 分析检测, 参数为默认设置。简单重复序列 (simple sequence repeats, SSRs) 利用 MISA (https://webblast.ipk-gatersleben.de/misa/) 进行分析, 参数设置为单核苷酸 (mononucleotide) 重复单元 ≥ 10 个, 二核苷酸 (dinucleotide) 重复单元 ≥ 5 个, 三核苷酸 (trinucleotide) 与四核苷酸 (tetranucleotide) 重复单元 ≥ 4 个, 五核苷酸 (pentanucleotide) 与六核苷酸 (hexanucleotide) 重复单元 ≥ 3 个, 且两个SSRs之间的距离不小于100 bp[30]。

密码子使用偏好性相关参数分析 为了避免短基因序列因密码子数量有限而导致密码子使用频率统计效力不足, 难以准确反映整体偏好性模式, 进而影响对自然选择或突变压力的解析[31], 从秦岭柴胡叶绿体全基因组序列中筛选长度大于 300 bp 的蛋白编码序列 (coding sequence, CDS) 用于密码子使用偏好性 (codon usage bias, CUB) 分析, 去除重复序列, 同时确保序列的起始密码子为ATG, 终止密码子为TAA、TGA、TAG, 共得到 52 条 CDS 序列。分别采用 CodonW[32] 软件和 EMBOSS[33] (https://www. bioinformatics. nl/emboss-explorer/) 中的 Cusp 和 Chips 工具对筛选得到的 52 条 CDS 序列的相对同义密码子使用度 (relative synonymous codon usage, RSCU)、密码子第 3 位上的碱基含量 (A_3、T_3、C_3、G_3) 和第 1、2、3 位碱基 GC 含量 (GC_1、GC_2、GC_3)、总 GC 含量 (GC_{all}) 及有效密码子数 (effective number of codon, ENC) 进行分析。

密码子使用偏好性影响因素分析 密码子使用偏好性是物种在长期演化过程中形成的遗传特征, 其形成受突变压力、自然选择、核苷酸组成等多种进化因素共同调控[34]。为探究秦岭柴胡叶绿体基因组序列密码子使用偏好性的可能影响因素, 对上述筛选得到的 52 条 CDS 序列分别进行中性绘图分析、ENC-plot 分析和 PR2-plot 分析。中性绘图分析以 GC_3 为横坐标、GC_{12} (GC_1 与 GC_2 的平均值) 为纵坐标, 绘制散点图并进行线性回归分析, 定量评估突变压力与自然选择的相对贡献。回归系数越接近 1, 表明密码子使用偏好性主要受基因突变影响; 而回归系数越接近 0, 则提示自然选择起主导作用[35]。ENC-plot 分析以 GC_3 为横坐标、ENC 为纵坐标, 使用 R 语言绘制散点图, 依据公式 ENC = 2+GC3s+29/[GC3s2+(1-GC3s)2] 绘制 ENC 期望值标准曲线, 通过量化密码子使用偏好强度 (ENC 值) 与密码子第三位 GC 含量 (GC3s) 的关联, 结合期望值标准曲线判断突变压力与自然选择的主导作用。当密码子使用偏好性主要受突变压力影响时, ENC 值分布于期望值标准曲线附近或其上; 当密码子使用偏好性主要受自然选择影响时, ENC 值偏离标准曲线[36]。密码子第 3 位碱基是影响其偏好性的关键位点, PR2-plot 分析以 $G_3/(G_3+C_3)$ 为横坐标、$A_3/(A_3+T_3)$ 为纵坐标, 绘制散点图, 分析密码子第 3 位碱基的使用差异[37]。

柴胡属种间叶绿体全基因组序列变异分析 目前柴胡属除秦岭柴胡外共有 39 个物种约 208 条 (截至 2025 年 8 月 31 日, 不包括缺少分类信息的物种) 叶绿体全基因组序列在 GenBank 数据库开放获取。为探究柴胡属植物叶绿体全基因组序列的种间变异情况, 下载这 39 个物种的叶绿体全基因组序列 (每个物种选取 1 条序列作为代表, 优先选取已正式发表个体), 使用 mVISTA 软件[38]基于 LAGAN 模型将秦岭柴胡与上述 39 个物种的叶绿体全基因组序列进行比对分析。同时, 利用在线工具 CPJSdraw[39] (http://112.86.217.82:9929/#/tool/alltool/detail/335) 分析各序列 IRs 区收缩/扩张情况。此外, 将包括秦岭柴胡在内的所有 40 个物种的叶绿体全基因组序列 (去掉一个 IR 区), 构建多序列矩阵, 使用 MAFFT 软件[40]完成序列比对, 并利用 DnaSP 软件[41]分析计算核苷酸多态性, 滑动窗口大小

设置为600 bp, 步长设置为200 bp[17]。

系统发育分析 将秦岭柴胡与上述39个物种的叶绿体基因组序列构建多序列比对矩阵, 以柴胡属近缘类群伞形科当归属*Angelica* L. 当归*A. sinensis* (Oliv.) Diels (MK688989) 和棱子芹属*Pleurospermum* Hoffmy. 岩生棱子芹*P. rupestre* (Popov) K. T. Fu & Y. C. Ho. (NC_071807) 作为外类群, 分别利用最大似然法 (maximum likelihood, ML) 和贝叶斯法 (Bayesian inference, BI) 进行系统发育重建分析。序列比对使用MAFFT完成, 并在Geneious中对比对结果进行校正。ML分析的最优核苷酸替换模型为GTR+G, 利用RAxML-HPC BlackBox v.8.1.4构建系统发育树, 靴带支持率 (bootstrap values, BS) 通过进行1 000次自展重复分析获得[42]。BI分析利用MrBayes v.3.2.7a[43]完成。通过JModeltest[44]在Akaike Information Criterion (AIC) 标准下获得最佳核苷酸替换模型TVM+I+G和相关参数, 以马尔科夫链的蒙特卡洛 (Markov Chain Monte Carlo, MCMC) 方法设置4条链同时运行, 从一棵随机选择的树开始运行10 000 000代 (ASDSF值小于0.01), 每1 000代保存一棵树, 舍弃25%的老化样本 (burn-in samples), 利用剩余的样本构建一致树 (consensus tree) 并计算后验概率 (posterior probability, PP)。ML和BI分析均在CIPRES[45] (https://www.phylo.org) 中完成。

结果

1 叶绿体全基因组测序、序列组装拼接与注释

组装拼接得到的秦岭柴胡完整环状四分体叶绿体全基因组见图1, 序列总长度为155 761 bp, 包括一个大单拷贝区 (large single copy, LSC; 85 663 bp)、一个小单拷贝区 (small single copy, SSC; 17 492 bp) 和一对反向重复区 (inverted repeats, IRa/IRb; 26 303 bp), 总GC含量为37.7%, LSC、SSC、IRs区和CDS序列的GC含量分别为35.8%、31.4%、42.8%和38.3% (表1)。

秦岭柴胡叶绿体全基因组序列共注释得到129个基因 (表2), 包括84个蛋白编码基因、37个tRNA基因和8个rRNA基因。其中, *rrn*16、*rrn*23、*rrn*4.5、*rrn*5、*trn*A-UGC、*trn*I-CAU、*trn*I-GAU、*trn*L-CAA、*trn*N-GUU、*trn*R-ACG、*trn*V-GAC和*rps*7、*rps*12、*rpl*2、*rpl*23、*ndh*B、*ycf*2共17个基因分别在两个反向重复区发生重复; *rps*19和*ycf*1基因分别跨LSC/IRb和SSC/IRa边界, 并在IRa和

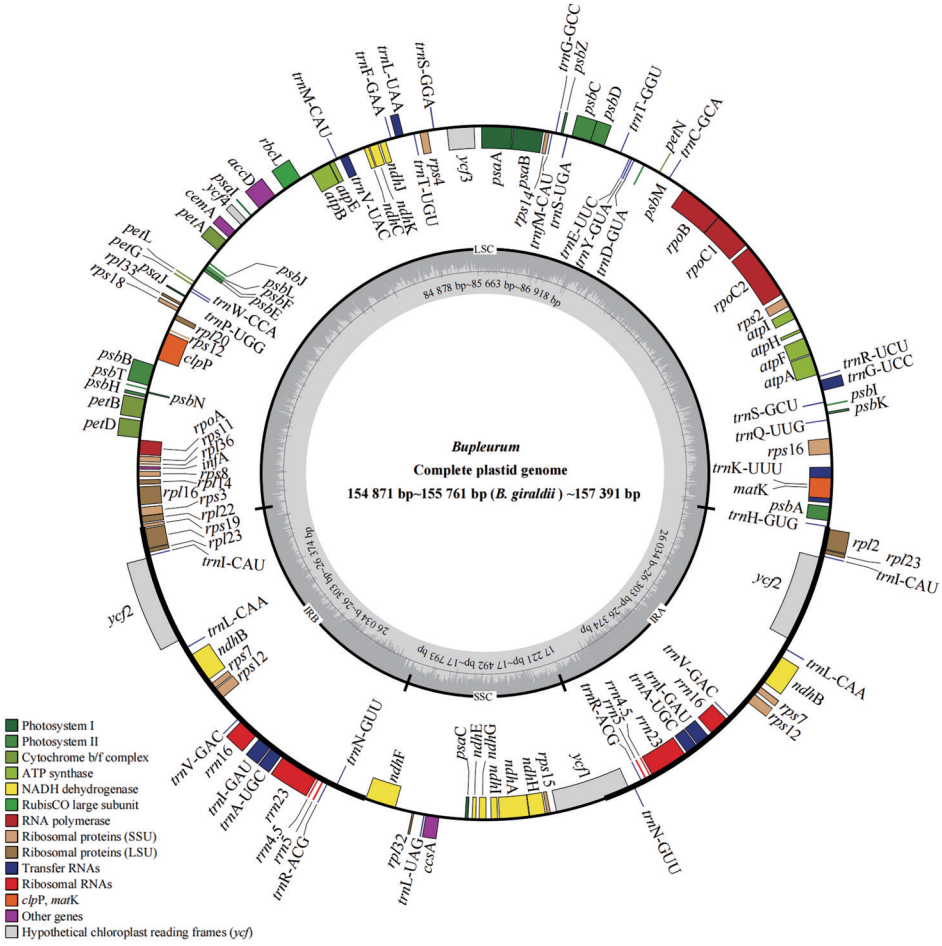


Figure 1 The complete plastid genome map of *B. giraldii* and other *Bupleurum* species

Table 1 Information of the plastid genomes of *B. giraldii* and the other 39 *Bupleurum* species. LSC: Large single copy; SSC: Small single copy; IRs: Inverted repeats; CDS: Coding sequence

Species	Accession	Genome		LSC		SSC		IRs		CDS	
		Length/bp	GC/%	Length/bp	GC/%	Length/bp	GC/%	Length/bp	GC/%	Length/bp	GC/%
B. giraldii	MW442884	155 761	37.7	85 663	35.8	17 492	31.4	26 303	42.8	79 805	38.1
B. angustissimum	NC_057639	155 829	37.7	85 598	35.8	17 599	31.3	26 316	42.9	79 815	38.1
B. aureum	NC_077617	155 568	37.7	85 428	35.9	17 510	31.4	26 315	42.8	79 872	38.1
B. bicaule	NC_057641	155 922	37.7	85 578	35.8	17 602	31.5	26 371	42.8	79 833	38.1
B. boissieuanum	MN756856	155 800	37.7	85 804	35.8	17 410	31.5	26 293	42.8	79 811	38.1
B. chinense	OR387523	155 790	37.7	85 729	35.8	17 501	31.4	26 280	42.8	79 894	38.1
B. candollei	MT261183	155 541	37.8	85 250	35.9	17 221	31.5	26 298	42.8	79 879	38.2
B. commelynoideum	OR508818	155 018	37.8	84 878	35.9	17 508	31.4	26 316	42.8	79 885	38.1
B. chaishoui	OQ460233	155 592	37.7	85 492	35.8	17 456	31.4	26 322	42.8	79 844	38.1
B. densiflorum	MT261184	155 787	37.7	85 473	35.8	17 592	31.4	26 361	42.7	79 870	38.1
B. euphorbioides	NC_053888	154 871	37.7	85 089	35.8	17 714	31.4	26 034	42.9	79 553	38.1
B. fruticosum	NC_060426	157 391	37.7	86 850	35.9	17 793	31.3	26 374	42.8	79 851	38.1
B. hamiltonii	NC_056298	155 320	37.8	85 280	35.9	17 468	31.6	26 286	42.8	79 802	38.2
B. kweichowense	NC_058630	155 025	37.7	84 922	35.9	17 505	31.4	26 311	42.8	79 870	38.1
B. kaoi	NC_063520	155 938	37.7	85 784	35.8	17 538	31.5	26 308	42.8	79 917	38.1
B. krylovianum	OQ621986	155 574	37.7	85 387	35.9	17 566	31.4	26 311	42.8	79 903	38.1
B. longiradiatum	OR508822	155 897	37.7	85 877	35.8	17 442	31.5	26 289	42.8	79 898	38.1
B. longicaule	OQ621987	155 149	37.8	85 114	35.9	17 419	31.7	26 308	42.8	79 932	38.2
B. marginatum var. *stenophyllum*	OR508825	155 553	37.8	85 353	35.9	17 602	31.7	26 299	42.8	79 882	38.2
B. marginatum	OR508826	155 478	37.7	85 329	35.9	17 605	31.7	26 272	42.8	79 882	38.1
B. malconense	OQ460229	155 562	37.7	85 476	35.9	17 468	31.4	26 309	42.8	79 849	38.1
B. microcephalum	OQ460231	155 080	37.8	85 319	36.0	17 511	31.6	26 125	42.9	79 846	38.2
B. pusillum	MT261188	155 224	37.7	85 132	35.8	17 504	31.4	26 294	42.8	79 861	38.1
B. petiolulatum	OQ621991	155 220	37.8	85 134	35.9	17 478	31.6	26 304	42.8	79 902	38.1
B. rockii	OQ621992	155 178	37.8	84 962	35.9	17 574	31.5	26 321	42.8	79 975	38.1
B. sibiricum	OR387525	155 777	37.7	85 537	35.8	17 610	31.2	26 315	42.8	79 900	38.1
B. sibiricum var. jeholense	OR387524	155 706	37.7	85 636	35.8	17 490	31.5	26 290	42.8	79 882	38.1
B. scorzonerifolium	OQ621993	155 962	37.6	85 676	35.8	17 618	31.2	26 334	42.8	79 888	38.1
B. smithii	OR387522	155 858	37.6	85 627	35.8	17 597	31.3	26 317	42.8	79 900	38.1
B. triradiatum	OQ621997	155 691	37.7	85 460	35.8	17 531	31.4	26 350	42.8	79 876	38.1
B. thianschanicum	OQ621995	155 385	37.7	85 054	35.9	17 611	31.4	26 360	42.7	79 847	38.1
B. wenchuanense	OQ621998	155 108	37.8	85 324	36.0	17 524	31.5	26 130	42.8	79 630	38.1
B. yunnanense	OQ621999	155 357	37.8	85 297	35.9	17 502	31.6	26 279	42.8	79 878	38.2
B. yinchowense	NC_054205	155 851	37.6	85 625	35.8	17 612	31.3	26 307	42.8	79 864	38.1
B. dracaenoides	NC_050984	155 140	37.8	85 323	35.9	17 537	31.6	26 140	42.8	79 831	38.1
B. falcatum	OR508821	155 872	37.7	85 742	35.8	17 528	31.4	26 301	42.8	79 901	38.1
B. gibraltaricum	NC_060427	157 303	37.7	86 918	35.8	17 643	31.3	26 371	42.8	79 826	38.1
B. latissimum	NC_033346	155 621	37.6	85 469	35.8	17 548	31.3	26 302	42.8	79 806	38.1
B. shanianum	NC_058629	154 925	37.8	84 925	35.9	17 478	31.6	26 261	42.8	79 787	38.2
B. sikangense	NC_056803	155 357	37.7	85 252	35.9	17 501	31.5	26 302	42.8	79 870	38.1

IRb区分别存在其对应长度的序列拷贝; *trn*A-UGC、*trn*G-UCC、*trn*I-GAU、*trn*K-UUU、*trn*L-UAA、*trn*V-UAC和*rpl*16、*rpl*2、*rpo*C1、*rps*16、*atp*F、*ndh*B、*ndh*A、*pet*B、*pet*D各包含有1个内含子, *clp*P、*ycf*3和*rps*12各包含2个内含子; *rps*12为反式剪切基因, 其5′端和3′端分别位于LSC区和IR区, 被分成2个独立的转录单元。

2 重复序列分析

柴胡属包括秦岭柴胡在内的40个物种 (各序列GenBank检索号见表1) 叶绿体全基因组LSC、SSC、IRs区及CDS序列中所包含的重复序列 (包括散在重复、串联重复和SSRs) 的分布信息见表3。秦岭柴胡叶绿体全基因组序列中共检测到31条长度≥30 bp的散在重复序列, 包括正向重复序列14条, 回文重复序列17条, 反向重复和互补重复序列均未检测到。其中, 有17条 (占比54.74%) 分别位于基因间区 (9条) 及内含子区 (8条)。检测到34条长度在9～49 bp, 重复次数在2～6次的串联重复序列, 有22条 (64.71%) 分别位于基因间区 (21条) 及内含子区 (1条)。检测到的SSRs的数量和类型见表4, 包括47个单核苷酸重复、9个双核苷酸重复、8个三核苷酸重复、1个五核苷酸重

Table 2 Gene composition in the plastid genome of *B. giraldii*. Bold font means genes duplicated in the IRs; a: Genes have two independent transcriptional units; b and c: Genes span the LSC/IRb and SSC/IRa regions, respectively; d and e: Genes have one and two introns, separately. NADH: Nicotinamide adenine diuncleotide; ATP: Adenosine triphosphate; CoA: Coenzyme A

Gene funcation	Type of genes	Name of genes
Self-replication	Ribosomal RNA	***rrn*4.5**, ***rrn*5**, ***rrn*16**, ***rrn*23**
	Transfer RNA	***trn*A-UGC**[d], *trn*C-GCA, *trn*D-GUC, *trn*E-UUC, *trn*F-GAA, *trnf*M-CAU, *trn*G-GCC, *trn*G-UCC[d], *trn*H-GUG, ***trn*I-CAU**, ***trn*I-GAU**[d], *trn*K-UUU[d], ***trn*L-CAA**, *trn*L-UAG, *trn*L-UAA[d], *trn*M-CAU, ***trn*N-GUU**, *trn*P-UGG, *trn*Q-UUG, ***trn*R-ACG**, *trn*R-UCU, *trn*S-GGA, *trn*S-UGA, *trn*S-GCU, *trn*T-UGU, *trn*T-GGU, ***trn*V-GAC**, *trn*V-UAC[d], *trn*W-CCA, *trn*Y-GUA
	Ribosomal protein (small subunit)	*rps*2, *rps*3, *rps*4, ***rps*7**, *rps*8, *rps*11, ***rps*12**[a,e], *rps*14, *rps*15, *rps*16[d], *rps*18, *rps*19[b]
	Ribosomal protein (large subunit)	***rpl*2**[d], *rpl*14, *rpl*16[d], *rpl*20, *rpl*22, ***rpl*23**, *rpl*32, *rpl*33, *rpl*36
	RAN polymerase	*rpo*A, *rpo*B, *rpo*C1[d], *rpo*C2
Photosynthesis	Subunits of NADH dehydrogenase	*ndh*A[d], ***ndh*B**[d], *ndh*C, *ndh*D, *ndh*E, *ndh*F, *ndh*G, *ndh*H, *ndh*I, *ndh*J, *ndh*K
	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*H, *psb*I, *psb*J, *psb*K, *psb*L, *psb*M, *psb*N, *psb*T, *psb*Z
	Subunits of cytochrome	*pet*A, *pet*B[d], *pet*D[d], *pet*G, *pet*L, *pet*N
	Subunits of ATP synthase	*atp*A, *atp*B, *atp*E, *atp*F[d], *atp*H, *atp*I
	Large subunit of rubisco	*rbc*L
Other genes	ATP-dependent protease	*clp*P[e]
	Maturase	*mat*K
	Envelope membrane protein	*cem*A
	Subunit of acetyl-CoA	*acc*D
	Synthesis gene	*ccs*A
	Proteins of unknown function	*ycf*1[c], ***ycf*2**, *ycf*3[e], *ycf*4, *inf*A

复和1个六核苷酸重复, 四核苷酸重复未检测到。其中, 单核苷酸A/T重复出现频率最高, 共45条, 占比68.18%, 其次为AT/AT (9条, 13.64%), 这些SSRs主要位于非编码的基因间区及内含子序列中 (55条, 83.33%), 并且主要分布在LSC区 (50个, 75.76%)。

3 密码子组成、使用偏好性相关参数及使用偏好性影响因素

秦岭柴胡叶绿体全基因组中筛选出的52条CDS序列共包含21 390个密码子, 编码20种氨基酸 (图2)。其中, 编码亮氨酸 (Leu) 的密码子使用频率最高, 共2 220次, 约占全部密码子总数的10.38%, 同义密码子为UUA、UUG、CUU、CUC、CUA、CUG; 编码半胱氨酸 (Cys) 的密码子使用率最低, 仅227次, 占全部密码子总数的1.06%, 同义密码子为UGU、UGC。RSCU值大于1的密码子有32个, 以A/U碱基结尾的有29个, 占比90.63%; RSCU值小于1的密码子有32个, 以A/U碱基结尾的密码子有3个, 占比9.38%。其中, 编码亮氨酸的密码子UUA的RSCU值最高, 为1.96, 编码甲硫氨酸 (Met) 的密码子AUG和编码色氨酸 (Trp) 的密码子UGG的RSCU值均为1, 编码丝氨酸 (Ser) 的密码子AGC的RSCU值最低, 为0.33。

52条蛋白编码基因序列密码子第1、2和3位碱基的GC含量GC_1、GC_2和GC_3依次分别为33.81%～58.40%、29.31%～56.12%和21.70%～37.84%, 平均GC含量呈现出GC_1 (46.98%) > GC_2 (39.70%) > GC_3 (28.98%), 且均小于50%, GC_{all}为30.45～46.52, 平均为38.55%。ENC值为36.71～55.36, 除*pet*D和*rpl*16外, 其余基因的ENC值均大于40。中性绘图分析结果显示, 所有基因均分布在对角线上方区域 (图3A), GC_3与GC_{12}的回归分析显示, 两者相关系数为0.053, 回归曲线斜率为0.325, GC_3与GC_{12}无显著相关性。PR2-plot分析结果显示 (图3B), 52条蛋白编码基因不均匀地分布在图中4个区域, 且主要集中在$G_3/(G_3+C_3) > 0.5$、$A_3/(A_3+T_3) < 0.5$的区域。ENC-plot分析结果显示多数叶绿体基因偏离ENC期望值标准曲线, 且多分布于曲线下方 (图3C)。

4 柴胡属叶绿体基因组序列种间变异比较分析

柴胡属包括秦岭柴胡在内的40个物种叶绿体全基因组序列所编码的基因类别、数量及编码顺序高度保守 (图4), 基因组总长度、LSC、SSC、IRs区及编码序列总长度存在较小差异, 其中叶绿体全基因组总长度变化范围为154 871 bp (*B. euphorbioides* Nakai)～157 391 bp (*B. fruticosum* L.), 相差2 520 bp, LSC、SSC、IRs区及编码序列长度最大差异依次分别为2 040、572、249和422 bp。序列的种间变异主要发生在非编码的基因间区, 且变异较高的区域多集中在LSC和SSC区, IRs区序列高度保守。所有40个物种叶绿体全基因组序列LSC、SSC及IRs区边界两侧的基因分布完全一致, IRs区没有出现明显扩张或收缩情况。其中, *rps*19基因跨LSC/IRb边界, 分别有195～

Table 3 Numbers and types of the tandem repeats (TRs), dispersed repeats (DRs) and simple sequence repeats (SSRs) detected in the LSC, SSC, and IRs of the plastid genomes of *B. giraldii* and the other 39 *Bupleurum* species

Species	LSC			SSC			IRs			CDS		
	TRs	DRs	SSRs	TRs	DRs	SSRs	TRs	DRs	SSRs	TRs	DRs	SSRs
B. giraldii	20	14	50	1	2	2	13	15	14	12	14	11
B. angustissimum	20	13	39	2	2	8	15	3	8	12	4	13
B. aureum	9	12	44	1	1	4	13	5	10	9	5	12
B. bicaule	24	13	38	2	2	10	17	4	12	12	5	13
B. boissieuanum	26	12	54	2	1	4	13	5	14	12	6	11
B. chinense	27	13	47	1	2	3	14	7	10	12	7	10
B. candollei	13	13	50	2	2	8	13	4	12	11	6	13
B. commelynoideum	16	12	35	1	1	4	13	6	12	12	5	10
B. chaishoui	18	11	44	2	2	5	13	5	12	11	6	12
B. densiflorum	12	14	45	2	2	6	17	4	14	11	5	12
B. euphorbioides	9	14	38	1	1	5	13	5	12	11	5	11
B. fruticosum	27	11	32	5	1	8	15	5	10	12	5	6
B. hamiltonii	16	12	48	2	2	7	15	5	14	11	6	12
B. kweichowense	13	12	39	1	2	6	13	5	12	11	6	11
B. kaoi	20	18	46	2	2	3	13	5	12	11	6	11
B. krylovianum	15	15	45	2	2	6	13	5	12	12	6	12
B. longiradiatum	27	19	43	2	2	4	13	5	10	12	6	10
B. longicaule	13	16	47	1	2	9	13	3	12	10	4	13
B. marginatum var. *stenophyllum*	9	12	44	1	2	8	17	5	18	11	6	12
B. marginatum	11	12	44	2	2	9	15	5	18	11	6	13
B. malconense	17	14	39	1	2	4	13	5	12	11	6	10
B. microcephalum	12	14	37	2	2	5	11	5	12	9	7	11
B. pusillum	13	13	41	1	2	4	13	4	14	11	6	10
B. petiolulatum	13	11	38	2	2	6	13	3	12	10	4	11
B. rockii	23	12	39	2	1	6	13	5	10	13	5	11
B. sibiricum	14	13	38	4	1	7	15	5	12	12	5	13
B. sibiricum var. jeholense	19	16	45	1	2	3	13	6	12	14	6	11
B. scorzonerifolium	19	13	38	3	2	7	17	6	12	13	6	12
B. smithii	22	13	37	2	1	8	15	5	12	13	5	12
B. triradiatum	14	18	45	3	2	5	17	5	14	11	5	11
B. thianschanicum	15	12	48	3	2	9	18	6	12	11	5	10
B. wenchuanense	11	14	45	2	2	7	11	5	12	10	6	11
B. yunnanense	13	12	49	3	2	8	13	3	12	10	4	12
B. yinchowense	20	14	39	4	1	7	13	5	12	10	5	12
B. dracaenoides	13	17	36	3	2	7	13	5	12	10	7	11
B. falcatum	21	18	44	1	2	5	13	6	14	11	6	10
B. gibraltaricum	29	10	36	6	1	8	15	5	8	11	5	11
B. latissimum	18	14	45	2	2	7	13	5	12	12	7	12
B. shanianum	11	11	46	2	2	10	15	5	16	11	6	13
B. sikangense	14	14	38	1	2	7	13	5	14	12	6	11

209 bp和70～84 bp的序列分布在LSC区和IRb区; *ycf*1跨SSC/IRa边界, 分别有3 582～3 661、1 851～1 902 bp的序列分布在SSC区和IRa区 (图5)。

核苷酸多样性分析结果显示 (图6), 秦岭柴胡及同属39个物种叶绿体全基因组序列的核苷酸多样性指数 (Pi) 为0.002 27～0.016 37, 平均值为0.009 86, LSC、SSC和IRs区序列的Pi值依次分别为0.001 57～0.033 59 (平均为0.010 95)、0.005 13～0.032 81 (平均为0.002 27) 和0～0.012 68 (平均为0.016 37), 共筛选到11条Pi值高于0.028 (0.028～0.034) 的种间高变异序列, 分别位于LSC区 (*trn*K_*rps*16、*trn*C_*psb*M、*pet*N_*psb*M、*pet*N_*psb*M+*psb*M) 和SSC区 (*ndh*F+*ndh*F_*rpl*32、*ndh*F_*rpl*32、*rpl*32+*rpl*32_*trn*L、*ycf*1×4)。

5 系统发育分析

系统发育分析结果显示 (图7), 柴胡属所有物种以BS_{ML} = 100%, PP = 1的支持率构成一个单系, 其中, 主要分布于地中海地区的*B. fruticosum* L.和*B. gibraltaricum* Lam.聚类为一支 (BS_{ML} = 100%, PP = 1), 位于最基部, 构成subgenus *Penninervias*; 其余所有物种均属于柴胡亚属, 以支持率BS_{ML} = 100%, PP = 1聚类为一个单系,

Table 4 Numbers and types of the detected SSRs in the plastid genome of *B. giraldii*

Repeat type	Type	Repeats	Number
Mononucletide	A	10	8
		11	6
		12	2
		13	2
		14	2
		19	1
	C	14	1
	G	14	1
	T	10	12
		11	5
		12	2
		13	2
		14	1
		15	1
		16	1
Dinucleotide	AT	5	1
		6	1
		7	1
		10	2
	TA	5	4
Trinucleotide	ATA	4	1
	ATT	4	2
	GAA	4	1
	GAT	4	1
	TAA	4	1
	TAT	4	1
	TTA	4	1
Pentanucleotide	ATAAT	3	1
Hexanucleotides	GGGAAT	3	1

并可进一步划分为4个次级单系分支 (clade I～IV, 除clade II的支持率为BS_{ML} = 60%, PP = 1外, 其他3个次级分支的支持率亦均为BS_{ML} = 100%, PP = 1), 秦岭柴胡位于clade I中, 和*B. falcatum* L.构成姐妹关系, 支持率亦为BS_{ML} = 100%, PP = 1。

讨论

秦岭柴胡在我国主要分布于陕西、山西、青海、甘肃及湖北等省份[46], 在陕西、甘肃及青海等地, 民间常以全草入药, 具有发表祛风、清肝利胆、清心火、通经之功效[10]。研究表明, 秦岭柴胡根及果实中含有丰富的柴胡总皂苷, 同时, 其茎、叶、果实中的黄酮类化合物含量均高于同属北柴胡、狭叶柴胡、小叶黑柴胡*B. smithii* Wolff var. *parvifolium* Shan et Y. Li等多个物种[47], 具有良好的药用价值和保护、研究及资源开发利用价值。本研究基于基因组浅层测序方法测序并组装分析获得秦岭柴胡叶绿体全基因组, 首次对其序列组成 (包括GC含量、基因分布模式、密码子使用偏好性及相对同义密码子使用度等) 与结构特征 (包括IRs区边界、重复序列分布等) 进行了全面解析, 并与同属已发表物种的叶绿体全基因组序列进行了比较分析。结果显示, 秦岭柴胡叶绿体基因组具有被子植物叶绿体基因组典型的环状四分体结构, 其序列总长度、CG含量、所编码的基因类别、数量及编码顺序与柴胡属其他已报道物种高度一致[17,19]。柴胡属包括秦岭柴胡在内的40个物种叶绿体全基因组序列总长度、LSC、SSC、IRs区及编码序列长度的种间差异范围依次为0～2 520 bp、3～2 040 bp、0～572 bp、0～249 bp和0～422 bp, 即种间序列长度差异主要来自LSC区。序列比对分析表明, LSC区长度的差异主要来自非编码序列中的重复序列 (包括散在重复序列、串联重复序列和SSRs)。秦岭柴胡LSC、SSC和IRs区的GC含量依次分别为35.8%、31.4%和42.8%, 即IRs区序列的GC含量显著高于LSC和SSC区。进一步分析显示, IRs区所包含的rRNA基因序列均具有较高的GC含量 (50.5%～56.5%, 平均为55.3%), 这可能是引起IRs区平均GC含量偏高的一个

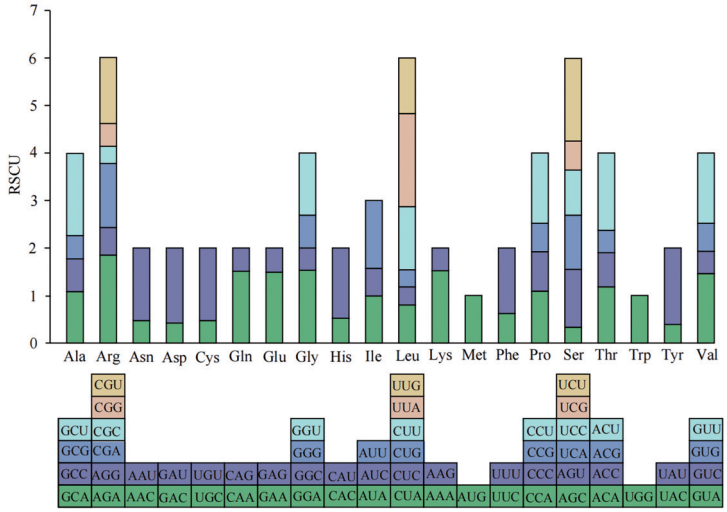


Figure 2 Codon contents of 20 amino acids in the 52 protein-coding genes in the plastid genome of *B. giraldii*. The top panel shows the relative synonymous codon usage (RSCU) for the corresponding amino acids

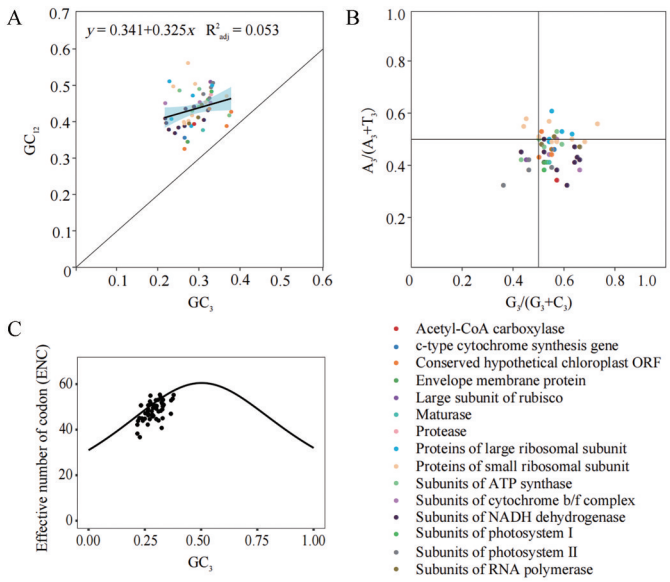


Figure 3 Codon usage bias analyses of the 52 protein-coding genes in the plastid genome *B. giraldii*. A: Neutral plot analysis; B: PR2-plot analysis; C: ENC-plot analysis. GC_{12}: Mean GC content at codon positions 1 and 2; A_3: A content at the third position of the codon; T_3: T content at the third position of the codon; G_3: G content at the third position of the codon; C_3: C content at the third position of the codon; GC_3: Codon third position GC content

重要因素。

序列变异分析结果表明, 柴胡属40个物种叶绿体全基因组序列的种间变异主要集中在非编码基因间区及内含子区, 且LSC区和SSC区序列的变异程度显著高于IRs区。柴胡属植物种间叶绿体全基因组序列LSC、SSC和IRs区的核苷酸多样性Pi值依次为LSC (0.01 095) > SSC (0.01 637) > IRs (0.00 227), 与序列变异分析结果一致。此外, 共筛选到11条种间高变异序列, 分别位于LSC区和SSC区, 且其中7条分布于基因间区, 进一步证明柴胡属植物叶绿体全基因组序列的种间变异主要发生在LSC和SSC区的非编码序列中, 与Zhang等[17]、Huang等[19]、Chao等[20]及Li等[48]对柴胡属的研究结果一致, 亦与重楼属 *Paris* L.[16]、黄精属 *Polygonatum* Mill.[49]及忍冬属 *Lonicera* L.[50]等被子植物其他类群叶绿体全基因组序列的种间变异模式一致。这11条种间高变异序列中, *trn*K_*rps*16、*pet*N_*psb*M、*ndh*F+*ndh*F_*rpl*32、*ycf*1及*rpl*32+*rpl*32_*trn*L分别与Zhang等[17]、Huang等[19]及Cheng等[49]对柴胡属、黄精属叶绿体基因组种间高变异序列分析筛选结果一致, 这些高变异序列可为柴胡属种间分子鉴定提供特异性候选分子标记。

简单重复序列是以少数核苷酸 (通常为1～6个) 为重复单元的短串联重复序列, 亦称微卫星DNA, 在真核生物基因组中广泛存在。因其多态性高、两端侧翼序列高度保守, 并且在叶绿体基因组中遵循单亲遗传模式, 作为分子标记在群体遗传学、进化生物学及分子育种等相关研究中被广泛使用[51]。本研究分析结果显示, 秦岭柴胡叶绿体全基因组序列中的SSRs以单核苷酸重复为主要类型 (占SSRs总数量的71.21%), 其次为二核苷酸重复 (13.64%), 且单核苷酸SSRs中, A/T碱基重复类型占比95.74%, 二核苷酸SSRs中AT/TA重复类型占比100%, 三核苷酸SSRs亦主要由A/T碱基组合 (ATA/ATT/TAA/TAT/TTA, 占比75%) 构成, 说明秦岭柴胡叶绿体基因组序列中的SSRs有明显的A/T碱基使用偏好。此外, 这些SSRs主要分布在LSC区 (50个, 占比75.76%), 两个IRs区分别检测到6个A/T碱基重复和1个G/C碱基重复, 而SSC区的2个SSRs亦均为A/T碱基重复。因此, 推测这些高A/T含量的SSRs可能是导致LSC区平均CG含量偏低的另一个原因。

遗传密码的简并性使得多数氨基酸能够被多个同义密码子编码。在演化过程中, 不同物种或同一物种的不同基因会表现出对特定同义密码子的非随机性选择, 即密码子使用偏好性。密码子使用偏好性在生物界中普遍存在, 其形成受自然选择与中性进化过程的共同作用[52,53]。对秦岭柴胡叶绿体基因组蛋白编码基因密码子使用偏好性分析结果表明, 当RSCU值大于1时, 密码子偏向以A/U碱基结尾。同时, 密码子三个位置的GC含量依次为GC_1 (33.81%～58.40%, 平均为46.98%) > GC_2 (29.31%～56.12%, 平均为39.70%) > GC_3 (21.70%～37.84%, 平均为28.98%), 且GC_{all}含量为30.45%～46.52% (平均为38.55%), 这些结果说明,

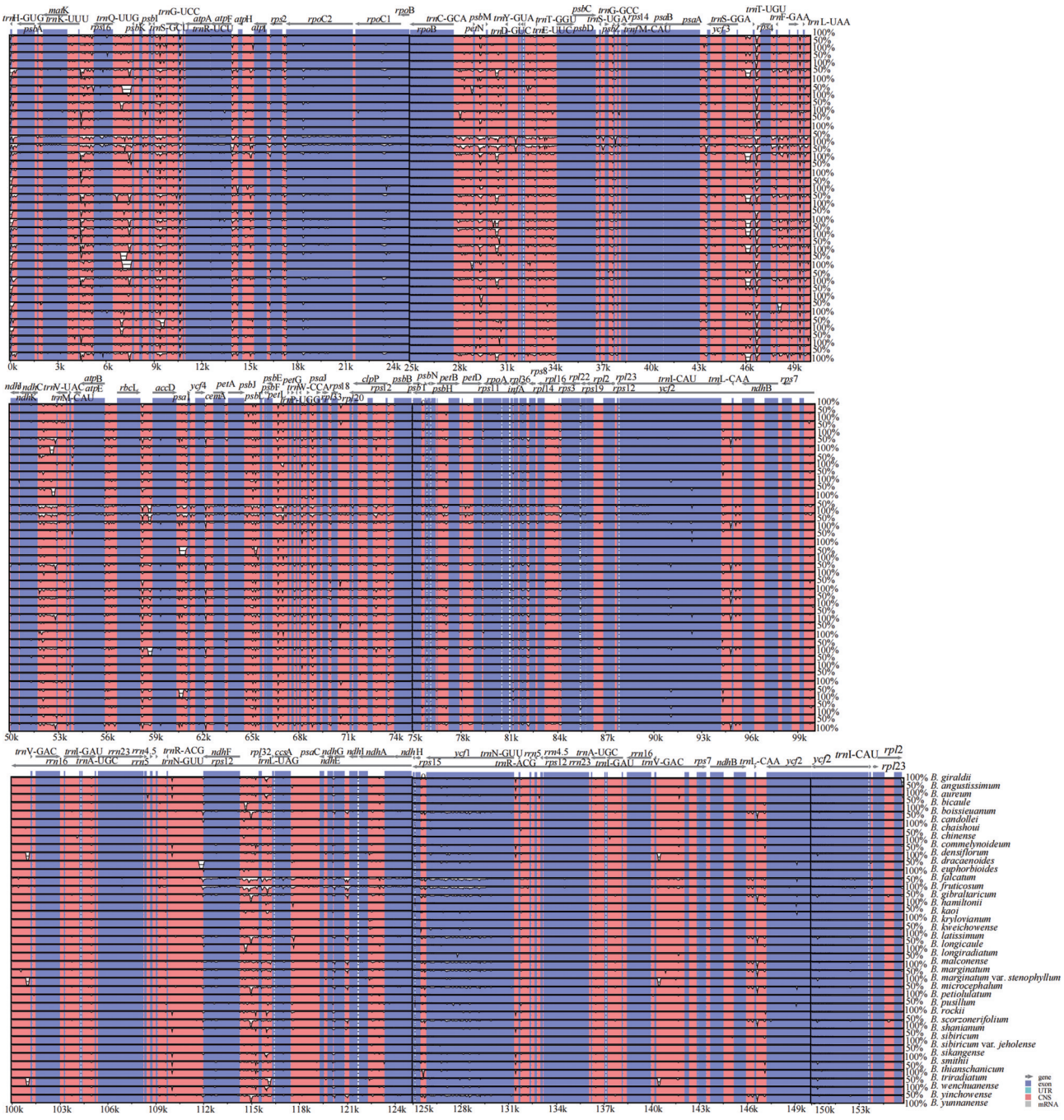


Figure 4 Sequences identity of the complete plastid genomes of *B. giraldii* and the other 39 *Bupleurum* species

秦岭柴胡叶绿体基因组在密码子不同位置上的GC含量分布不均匀, 整体上表现出对以A/U结尾的密码子的使用偏好。这也与主要分布于非编码区的SSRs的碱基组成A/T偏好呈现出高度统一性, 表明秦岭柴胡叶绿体基因组编码区的密码子优化与非编码区的重复序列演化受到相似的碱基组成约束与演化驱动力, 进一步反映出细胞器基因组在结构与功能层面的协同演化规律。中性绘图分析结果显示, 秦岭柴胡叶绿体基因组中52个蛋白编码基因分布相对集中, 均位于对角线上方, 且GC_{12}与GC_3无显著相关性 (回归曲线相关系数为0.053), 说明秦岭柴胡叶绿体基因组蛋白编码基因密码子的使用偏好性主要受自然选择的影响[54]。ENC值是衡量密码子使用偏好性强弱的重要指标, 秦岭柴胡叶绿体基因组蛋白编码基因ENC值为36.71～55.36, 除*pet*D和*rpl*16外, 其余基因的ENC值均大于40, 表明整体密码子使用偏好性较弱。此外, ENC-plot分析结果显示多数基因分布点偏离ENC期望值标准曲线, 且普遍分布于标准曲线下方, 说明密码子使用偏

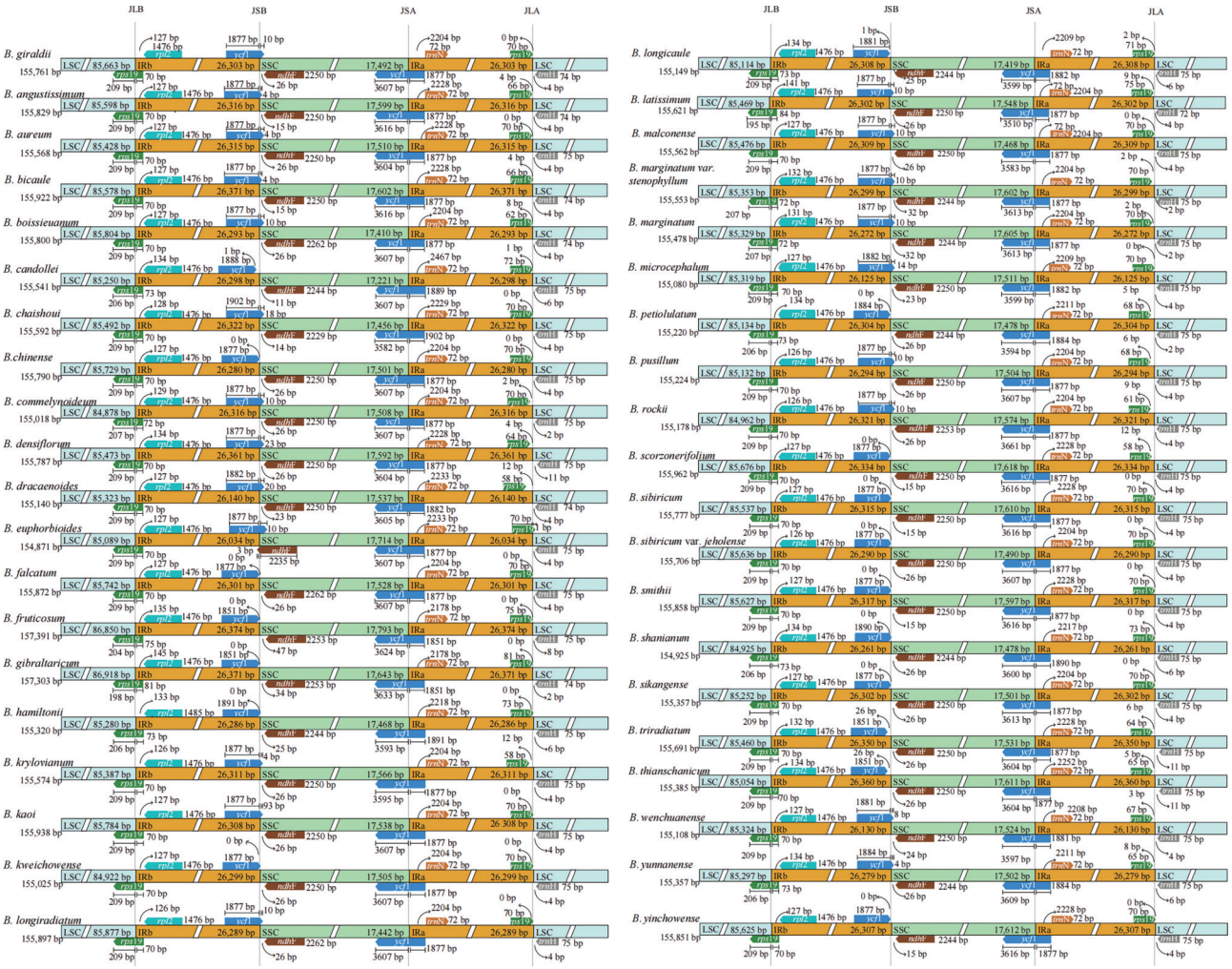


Figure 5 Comparison of the borders of LSC, SSC and IRa/b regions of the complete plastid genomes of *B. giraldii* and the other 39 *Bupleurum* species

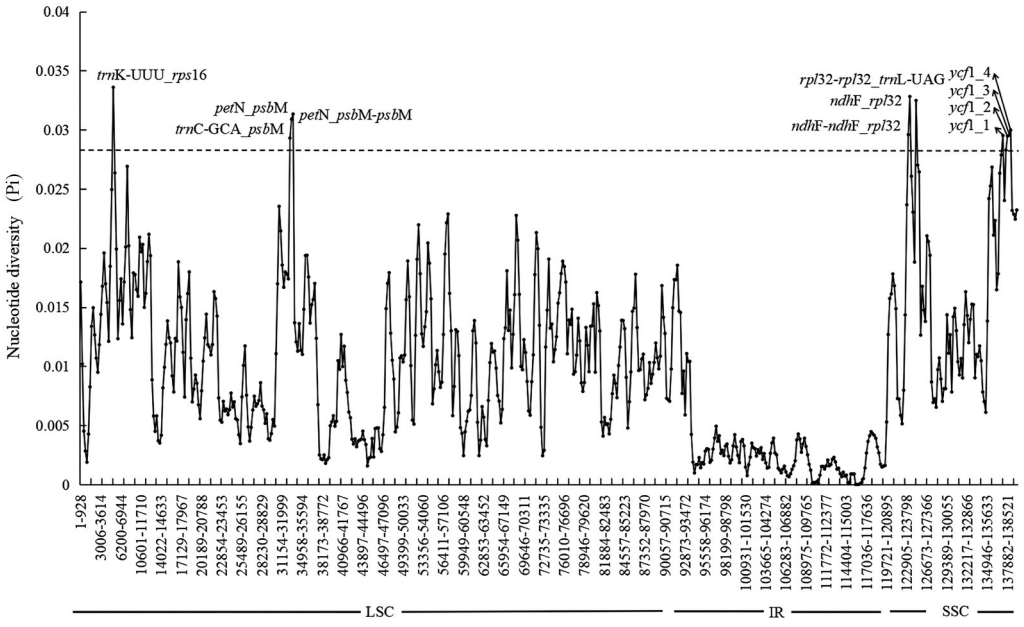


Figure 6 The nucleotide diversity of the complete plastid genome sequences of *B. giraldii* and the other 39 *Bupleurum* species

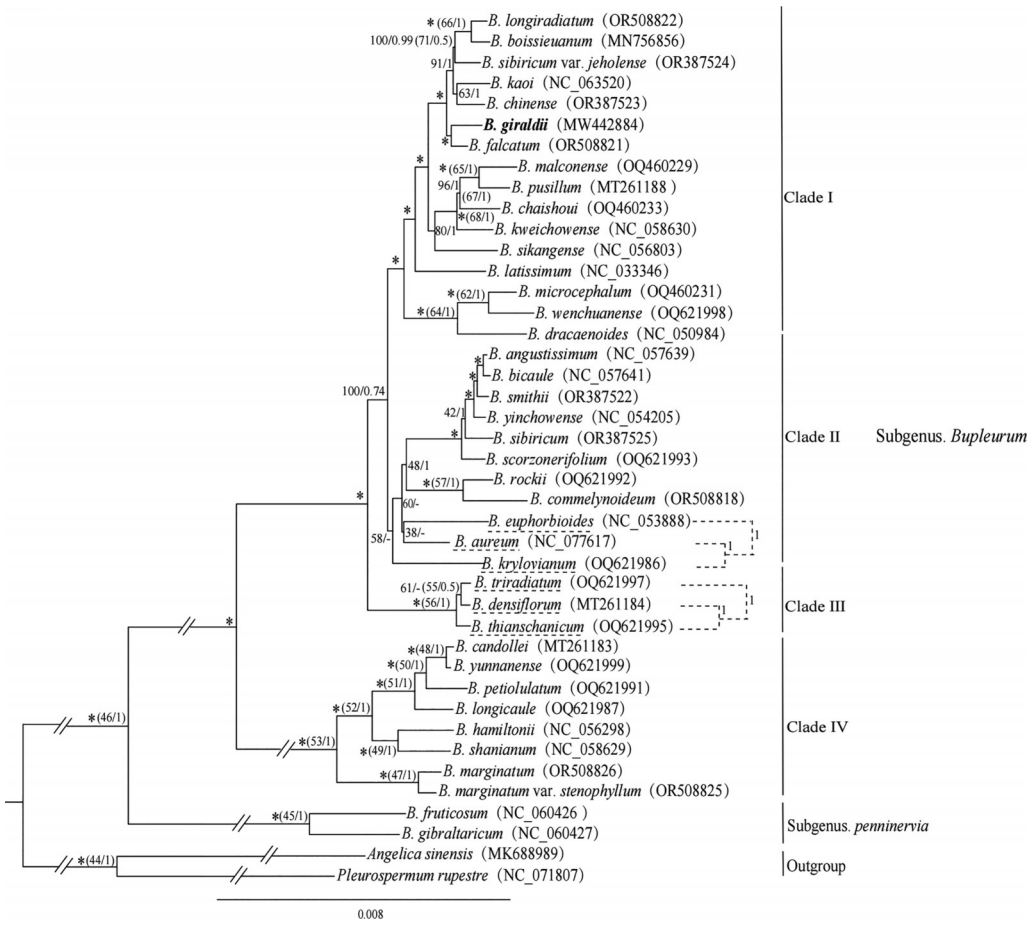


Figure 7 Phylogenetic tree of *Bupleurum* species constructed using the complete plastid genome sequences and the concatenated 11 highly divergent sequences based on Maximum likelihood (ML) and Bayesian inference (BI) methods (differences in topology were indicated by dotted lines). Numbers beside each node are bootstrap support values (BS) from maximum likelihood and posterior probabilities (PP) from Bayesian inference. * indicates nodes with BS_{ML} = 100%, PP = 1; numbers in brackets indicate BS and PP from the concatenated 11 highly divergent sequences

好受自然选择主导[55]。PR2-plot分析结果显示, 52个蛋白编码基因集中分布在$G_3/(G_3+C_3) > 0.5$、$A_3/(A_3+T_3) < 0.5$的区域, 说明秦岭柴胡叶绿体基因组密码子第三位的碱基使用频率T > A、G > C, 即存在显著的密码子使用偏好性。结合中性绘图与ENC-plot等分析结果, 进一步表明自然选择作用是驱动秦岭柴胡叶绿体基因组密码子使用偏好性形成的主导因素。

Song等[5]、Wang等[8]基于ITS、*trn*H-psbA、*mat*K及*rbc*L序列对柴胡属的分子系统发育分析结果将秦岭柴胡置于柴胡亚属, 并与贵州柴胡或贵州柴胡+空心柴胡亲缘关系最近, 但支持率普遍偏低 (BS_{ML} = 67%～78%, PP = 0.75～0.95)。本研究在增加物种取样数量的基础上基于叶绿体全基因组序列的系统发育分析结果以BS_{ML} = 100%, PP = 1的支持率支持柴胡属所有物种构成一个单系, 属下划分为两个单系, 分别对应柴胡亚属和*Penninervias*亚属, 秦岭柴胡位于柴胡亚属, 并与*B. falcatum* L.构成姐妹关系, 支持率亦为BS_{ML} = 100%, PP = 1。同时, 系统树中大多数分支节点的支持率均为BS_{ML} = 100%, PP = 1, 且种间系统发育关系与其他基于叶绿体全基因组序列的系统发育重建分析结果高度一致[19,20,56]。为了进一步探讨种间高变异序列对柴胡属种间系统发育关系的分辨率, 将本研究筛选的11条种间高变异序列分别提取出来, 构建联合矩阵, 以最大似然法和贝叶斯法进行系统发育重建分析, 结果显示, 所得系统发育树拓扑结构与基于叶绿体全基因组序列的分析结果基本一致, 但节点支持率显著降低 (BS_{ML} = 44%～84%, PP = 0.5～1), 说明相较于单个或有限的基因片段, 叶绿体全基因组包含更丰富的物种演化信息, 是解析柴胡属种间系统发育关系问题的一个有效途径。

形态上, 秦岭柴胡植株通常较矮小, 以茎丛生、基生叶渐狭成长叶柄为显著识别特征。Wolff等[57]首次

将其发表为长茎柴胡的变种, 即 *Bupleurum longicaule* Wall. ex DC. var. *giraldii* H. Wolff。此后, Shan 等[58]和 She 等[1]在《中国植物志》及 Flora of China 中均沿用了这一变种处理, 将其置于长茎柴胡复合群。Wang 等[8]综合形态与分子系统学的分析结果发现, 秦岭柴胡与长茎柴胡原变种 *B. longicaule* var. *longicaule* 的亲缘关系较远, 在系统发育树上分属不同的支系; 结合其茎丛生、叶具长柄等稳定的形态差异, 遂将其提升为独立的种 *Bupleurum giraldii* (H. Wolff) Koso-Pol.。同时, 该研究结果显示, 空心柴胡与贵州柴胡以高支持率 (BS_{ML} = 96%, PP = 1) 聚为一支 (东亚 Group II 内部的 Clade E), 秦岭柴胡虽也属于东亚 Group II, 但分化时间较早, 系统发育位置相对独立。*Bupleurum falcatum* L. 由林奈于 1753 年在《植物种志》(Species Plantarum) 中正式描述并命名。该物种茎通常单生或少数丛生, 直立, 以叶片常呈镰刀状弯曲为典型特征。由于其地理分布范围横跨欧洲至中亚, 不同地理居群间常存在一定程度的形态变异。Song 等[5]的研究指出, 秦岭柴胡和 *B. falcatum* 均起源于地中海盆地, 后经中亚扩散至东亚并独立演化。此外, 两者均以茎实心而与空心柴胡及同属其他物种相区别, 从地理分布和形态特征上为两者的近缘关系提供了支持。然而, 由于本研究仅以单一个体代表各物种, 难以全面揭示种内遗传变异, 后续研究需从物种水平与不同地理分布居群层面扩大取样, 结合地理分布、形态性状、核基因组数据等多方面证据, 进一步系统解析柴胡属种间划分和系统演化关系问题。

作者贡献: 共同第一作者陈锦豪、赵晨负责实验、数据分析和论文初稿撰写; 通讯作者张明英负责实验设计、数据分析和论文指导; 高静、李依民、陈莹参与实验及数据分析; 张岗、颜永刚参与实验材料采集与论文指导。所有作者参与论文修改。

利益冲突: 本文所有作者之间不存在利益冲突。

References

[1] She ML, Watson MF. Apiaceae (Umbelliferae) in Flora of China [M]. Beijing: Science Press, 2005, 14: 60-74.

[2] Li SJ. *Bupleurum* L. in Medicinal Flora of China (中国药用植物志) [M]. Beijing: Peking University Medical Press, 2018, 7: 607-647.

[3] Sun P, Li YJ, Wei S, et al. Pharmacological effects and chemical constituents of *Bupleurum* [J]. Mini Rev Med Chem, 2019, 19: 34-55.

[4] Yu X, Miao Z, Zhang L, et al. Extraction, purification, structure characteristics, biological activities and pharmaceutical application of Bupleuri Radix polysaccharide: A review [J]. Int J Biol Macromol, 2023, 237: 124146.

[5] Song XY, Yang YC, Zhou YY, et al. Speciation and diversification of the *Bupleurum* (Apiaceae) in East Asia [J]. PhytoKeys, 2024, 248: 41-57.

[6] Shu P, Yuan CQ, She ML, et al. Numerical taxonomy of medicinal *Bupleurum* species in China (I) [J]. Acta Bot Boreali Occident Sin (西北植物学报), 1998, 18: 277-283.

[7] Neves SS, Watson MF. Phylogenetic relationships in *Bupleurum* (Apiaceae) based on nuclear ribosomal DNA ITS sequence data [J]. Ann Bot, 2004, 93: 379-398.

[8] Wang CB, Ma XG, He XJ. A taxonomic re-assessment in the Chinese *Bupleurum* (Apiaceae): Insights from morphology, nuclear ribosomal internal transcribed spacer, and chloroplast (*trn*H-*psb*A, *mat*K) sequences [J]. J Syst Evol, 2011, 49: 558-589.

[9] Northwest Institute of Botany, Chinese Academy of Sciences. Flora of Qinling Mountains (秦岭植物志) [M]. Vol 1. Fasc 3. Beijing: Science Press, 1981: 392.

[10] Wei H, Liu Q, Wei WG. Experimental study on dose-effect relationship of hepatoprotective effect of the different extraction from *Bupleurum longicaule* var. *giraldii* [J]. Chin J Exp Tradit Med Formulae (中国实验方剂学杂志), 2012, 18: 237-240.

[11] Bobik K, Burch-Smith TM. Chloroplast signaling within, between and beyond cells [J]. Front Plant Sci, 2015, 6: 781.

[12] Hagemann R. The sexual inheritance of plant organelles [M]// Daniell H, Chase C D. Molecular Biology and Biotechnology of Plant Organelles. Netherlands: Springer, 2004: 93-113.

[13] Chen HY, Zhang ZR, Yao X, et al. Plastid phylogenomics provides new insights into the systematics, diversification, and biogeography of *Cymbidium* (Orchidaceae) [J]. Plant Divers, 2024, 46: 448-461.

[14] Liao X, Mahai R, Wang X, et al. Characterization and phylogenomic inference of chloroplasts in seven Apioideae species [J]. Genet Resour Crop Evol, 2026, 73: 50.

[15] Wang T, Feng H, Zhu H, et al. Molecular phylogeny and comparative chloroplast genome analysis of the type species *Crucigenia quadrata* [J]. BMC Plant Biol, 2025, 25: 64.

[16] Miao K, Wang T, Tang L, et al. Establishing the first reference library for utilizing high-throughput sequencing technologies in identifying medicinal and endangered *Paris* species (Melanthiaceae) [J]. Ind Crops Prod, 2024, 218: 118871.

[17] Zhang MY, Zhang YQ, Li YM, et al. Complete plastid genomes of *Bupleurum chinense* DC. and *B. boissieuanum* H. Wolff, with comparative and phylogenetic analyses of medicinal *Bupleurum* species [J]. Acta Pharm Sin (药学学报), 2021, 56: 618-629.

[18] Yang L, Xiong F, Xiao Y, et al. The complete chloroplast genome of *Bupleurum longicaule* var. *strictum*, an annual herb endemic to China [J]. Mitochondrial DNA B Resour, 2020, 5: 899-901.

[19] Huang R, Xie X, Chen A, et al. The chloroplast genomes of four *Bupleurum* (Apiaceae) species endemic to southwestern China,

a diversity center of the genus, as well as their evolutionary implications and phylogenetic inferences [J]. BMC Genomics, 2021, 22: 714.

[20] Chao Z, Hong X, Xie X, et al. Phylogenomic analysis of *Bupleurum* in western Sichuan, China, including an overlooked new species [J]. Front Plant Sci, 2023, 14: 1294670.

[21] Patel RK, Jain M. NGS QC Toolkit: A toolkit for quality control of next generation sequencing data [J]. PLoS One, 2012, 7: e30619.

[22] Jin JJ, Yu WB, Yang JB, et al. GetOrganelle: A fast and versatile toolkit for accurate *de novo* assembly of organelle genomes [J]. Genome Biol, 2020, 21: 241.

[23] Wick RR, Schultz MB, Zobel J, et al. Bandage: Interactive visualization of *de novo* genome assemblies [J]. Bioinformatics, 2015, 31: 3350-3352.

[24] Kearse M, Moir R, Wilson A, et al. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data [J]. Bioinformatics, 2012, 28: 1647-1649.

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

[26] Chan PP, Lowe TM. tRNAscan-SE: Searching for tRNA genes in genomic gequences [J]. Methods Mol Biol, 2019, 1962: 1-14.

[27] Greiner S, Lehwark P, Boc R. Organellar GenomeDRAW (OGDRAW) version 1.3.1: Expanded toolkit for the graphical visualization of organellar genomes [J]. Nucleic Acids Res, 2019, 47: W59-W64.

[28] 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.

[29] Benson G. Tandem repeats finder: A program to analyze DNA sequences [J]. Nucleic Acids Res, 1999, 27: 573-580.

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

[31] Mu JJ, Zhang JS. Codon usage bias analysis in the chloroplast genome of *Actinostemma tenerum* (Cucurbitaceae) [J]. Curr Issues Mol Biol, 2025, 47: 833.

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

[33] Rice P, Longden I, Bleasby A. EMBOSS: The European molecular biology open software suite [J]. Trends Genet, 2000, 16: 276-277.

[34] Gao MQ, Huo XW, Lu LT, et al. Analysis of codon usage patterns in *Bupleurum falcatum* chloroplast genome [J]. Chin Herb Med, 2023, 15: 284-290.

[35] Sueoka N. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons [J]. J Mol Evol, 2004, 40: 318-325.

[36] Wright F. The 'effective number of codons' used in a gene [J]. Gene, 1990, 87: 23-29.

[37] Romero H, Zavala A, Musto H. Codon usage in chlamydia trachomatis is the result of strand-specific mutational biases and a complex pattern of selective forces [J]. Nucleic Acids Res, 2000, 28: 2084-2090.

[38] Frazer KA, Pachter L, Poliakov A, et al. VISTA: Computational tools for comparative genomics [J]. Nucleic Acids Res, 2004, 32: W273-W279.

[39] Huie L, Qiqiang G, Lei X, et al. CPJSdraw: Analysis and visualization of junction sites of chloroplast genomes [J]. Peer J, 2023, 11: e15326.

[40] Katoh K, Toh H. Parallelization of the MAFFT multiple sequence alignment program [J]. Bioinformatics, 2010, 26: 1899-1900.

[41] Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, et al. DnaSP 6: DNA sequence polymorphism analysis of large data sets [J]. Mol Biol Evol, 2017, 34: 3299-3302.

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

[43] Ronquist F, Teslenko M, Van der Mark P, et al. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space [J]. Syst Biol, 2012, 61: 539-542.

[44] Darriba D, Taboada GL, Doallo R, et al. jModelTest 2: More models, new heuristics and high-performance computing [J]. Nat Methods, 2012, 9: 772.

[45] Miller MA, Pfeiffer W, Schwartz T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees [C]. 2010 Gateway Computing Environments Workshop (GCE). IEEE, 2010: 1-8.

[46] Editorial Board of Chinese Flora, Chinese Academy of Sciences. Liliaceae in Flora Reipublaris Sinicae (中国植物志) [M]. Vol 55. Beijing: Science Press, 1979: 251.

[47] Zhou YF. Studies on the Comparision of Structures and Their Accumulation of Chemical Components and the Development of Secretory Canals in Six Medicinal Plants of *Bupleurum* L. (柴胡属6种药用植物结构与化学成分积累的比较和分泌道形态发生的研究) [D]. Xi'an: Northwest University, 2008.

[48] Li J, Xie DF, Guo XL, et al. Comparative analysis of the complete plastid genome of five *Bupleurum* species and new insights into DNA barcoding and phylogenetic relationship [J]. Plants, 2020, 9: 543.

[49] Cheng WP, Zhao X, Li YM, et al. Characterization of the complete chloroplast genome of *Polygonatum franchetii* Hua, a Chinese endemic medicinal species, and phylogenetic relationships of *Polygonatum* [J]. Acta Pharm Sin (药学学报), 2023, 58: 3461-3472.

[50] Zhang JJ, Yuan Q, Wei Y, et al. Evolutionary analysis of chloro-

plast genomes of Sect. *Isika* (*Lonicera*) species [J]. Chin Tradit Herb Drugs (中草药), 2024, 55: 3085-3097.

[51] Ellegren H. Microsatellites: Simple sequences with complex evolution [J]. Nat Rev Genet, 2004, 5: 435-445.

[52] Powell W, Morgante M, McDevitt R, et al. Polymorphic simple sequence repeat regions in chloroplast genomes: applications to the population genetics of pines [J]. Proc Natl Acad Sci U S A, 1995, 92: 7759-7763.

[53] Buchan JR, Aucott LS, Stansfield I. tRNA properties help shape codon pair preferences in open reading frames [J]. Nucleic Acids Res, 2006, 34: 1015-1027.

[54] Song H, Liu J, Song Q, et al. Comprehensive analysis of codon usage bias in seven *Epichloë* species and their peramine-coding genes [J]. Front Microbiol, 2017, 8: 1419.

[55] Lü H, Zhao WM, Zheng Y, et al. Analysis of synonymous codon usage bias in *Chlamydia* [J]. Acta Biochim Biophys Sin, 2005, 37: 1-10.

[56] Wang P, Bai J, Li X, et al. Phylogenetic relationship and comparative analysis of the main Bupleuri Radix species in China [J]. PeerJ, 2023, 11: e15157.

[57] Wolff H. Umbelliferae-Apioideae-*Bupleurum*, Trinia et reliquae Ammineae heteroclitae [M] // Engler A. Das Pflanzenreich: *Regni vegetabilis conspectus* (植物界: 植物界概览). Heft 43 (IV. 228). Leipzig: Engelmann, 1910.

[58] Shan RH, Li Y. On the Chinese species of *Bupleurum* L. [J]. J Univ Chin Acad Sci (中国科学院大学学报), 1974, 12: 261-295.