曾文娟,周 伟,何诗恬,等. 茶树品种白毫早叶绿体基因组结构特征及其密码子偏好性分析[J]. 江苏农业学报,2025,41(7):1398-1411.
doi:10.3969/j.issn.1000-4440.2025.07.016

茶树品种白毫早叶绿体基因组结构特征及其密码子偏好性分析

曾文娟[1,2,3,4], 周 伟[1,2,3,4], 何诗恬[1,2,3,4], 贺 宁[1,2,3,4], 龚意辉[1,2,3,4], 陈致印[1,2,3,4]
(1.湖南人文科技学院农业与生物技术学院,湖南 娄底 417000; 2.湖南人文科技学院湖南省园艺生产与加工类创新创业教育中心,湖南 娄底 417000; 3.湖南人文科技学院湖南省创新创业示范基地,湖南 娄底 417000; 4.湖南人文科技学院湖南省湘中特色农业资源开发利用与质量安全控制重点实验室,湖南 娄底 417000)

摘要: 本研究首次解析了茶树品种白毫早(*Camellia sinensis* cv. Baihaozao)的完整叶绿体基因组特征。采用 Illumina 高通量测序技术对白毫早的叶绿体基因组进行从头组装,并借助 Geneious 平台进行基因注释及微卫星标记检测。同时基于 IRscope 工具实现基因组结构的可视化分析。在密码子使用偏好性评估中,综合采用相对同义密码子使用度(*RSCU*)、有效密码子数(*ENC*)及中性绘图法系统解析密码子选择模式及其演化驱动力。结果表明,白毫早茶树叶绿体基因组是总长度为157 025 bp 的典型四分体环状结构,含有大小为86 586 bp 的大单拷贝区(LSC)、18 277 bp 的小单拷贝区(SSC)及 1 对大小各为26 081 bp 的反向重复序列(IR)区域;全基因组的 G+C 含量为 37.30%,其中 IR 区域的G+C 含量最高(42.95%),SSC 区域的 G+C 含量最低(30.55%)。白毫早的叶绿体环状基因组共注释了 133 个功能基因,包括 87 个蛋白质编码基因、37 个转运核糖核酸(tRNA)基因、8 个核糖体核糖核酸(rRNA)基因及 1 个假基因,基因中内含子与外显子的分布呈现显著差异,如 *trnK-UUU* 的内含子最大,为2 488 bp,*ycf2* 的外显子最长,为6 897 bp。此外,本研究检测到 247 个简单重复序列(SSR)位点,其中单核苷酸重复占比较高(占比为 63.56%),且表现出明显的 A/T 偏好性(97.45%);本研究鉴定出 40 个长重复序列,包含 20 个正向重复序列、20 个回文重复序列。密码子偏好性分析结果显示,密码子第 3 位碱基中 G+C 的平均含量(GC_3值)(27.59%)显著低于密码子第 1 位碱基中 G+C 的平均含量(GC_1值)(46.85%)、密码子第 2 位碱基中 G+C 的平均含量(GC_2)(39.50%)。*ENC* 的均值为 44.57,表明密码子的偏好性较弱。中性回归分析结果显示,校正决定系数(R^2_{adj})= 0.016 0,自然选择是起主导作用的进化动力(贡献率为 91.47%)。90%最优密码子以 A/T 结尾,这一特征与通过奇偶规则 2(PR2)分析得出的密码子第 3 位的 G 含量(G_3)>密码子第 3 位碱基中的 C 含量(C_3)、密码子第 3 位的 T 含量(T_3)>密码子第 3 位的 A 含量(A_3)的偏好性结果一致。基于 31 个保守蛋白质编码基因构建系统发育树,结果显示,白毫早茶树与福鼎白毫茶树形成高支持的分支[自举值(*bootstrap*)= 100%],属于山茶属核心类群。本研究结果为茶树遗传资源评价及叶绿体基因组进化机制的研究提供了参考。

关键词: 茶树; 叶绿体基因组; 密码子偏好性; 最优密码子
中图分类号: S571.1 **文献标识码:** A **文章编号:** 1000-4440(2025)07-1398-14

Structural characteristics and codon usage bias analysis of the chloroplast genome in the tea cultivar *Camellia sinensis* cv. Baihaozao

ZENG Wenjuan[1,2,3,4], ZHOU Wei[1,2,3,4], HE Shitian[1,2,3,4], HE Ning[1,2,3,4], GONG Yihui[1,2,3,4], CHEN Zhiyin[1,2,3,4]
(1. *College of Agriculture and Biotechnology, Hunan University of Humanities, Science and Technology, Loudi* 417000, *China*; 2.*Innovation and Entrepreneurship Education Center for Horticultural Production and Processing in Hunan Province, Hunan University of Humanities, Science and Technology, Loudi* 417000, *China*; 3.*Hunan University of Humanities, Science and Technology, Hunan Provincial Innovation and Entrepreneurship Demonstration Base, Loudi* 417000; 4.*Key Laboratory of Characteristic Agricultural Resource Development and Quality Safety Con-*

收稿日期:2025-04-28
基金项目:湖南省自然科学基金项目(2023JJ50465);湖南省科技创新计划项目(2024RC8289);娄底市科技创新计划项目(2023-RC3501);国家级大学生创新训练项目(S202310553022)
作者简介:曾文娟(2004-),女,湖南邵阳人,本科生,主要从事茶学研究。(E-mail)2103562420@qq.com。周伟为共同第一作者。
通讯作者:陈致印,(E-mail)772612626@qq.com

trol in Hunan Province, *Hunan University of Humanities*, *Science and Technology*, *Loudi* 417000, *China*)

Abstract: This study reports the first complete chloroplast genome of the tea cultivar *Camellia sinensis* cv. Baihaozao. Using Illumina high-throughput sequencing, we *de novo* assembled the chloroplast genome and performed gene annotation and microsatellite detection in Geneious. Additionally, genome structure visualization analysis was achieved with the IRscope tool. In the assessment of codon usage bias, we comprehensively employed the relative synonymous codon usage (*RSCU*), the effective number of codons (*ENC*), and neutrality plot analysis to systematically analyze codon selection patterns and their evolutionary driving forces. The results revealed that the chloroplast genome of *Camellia sinensis* cv. Baihaozao exhibited a typical quadripartite circular structure with a total length of 157 025 bp, comprising a large single-copy (LSC) region of 86 586 bp, a small single-copy (SSC) region of 18 277 bp, and a pair of inverted repeat (IR) regions each 26 081 bp in length. The overall G+C content of the chloroplast genome was 37.30%, with the IR regions exhibiting the highest value (42.95%) and the SSC region showing the lowest (30.55%). The circular chloroplast genome was annotated with 133 functional genes, comprising 87 protein-coding genes, 37 transfer RNA (tRNA) genes, eight ribosomal RNA (rRNA) genes, and one pseudogene. Gene structure analysis revealed marked variation in intron and exon distribution. The largest intron (2 488 bp) was identified in *trnK-UUU*, while the longest exon (6 897 bp) was found in *ycf2*. Furthermore, 247 simple sequence repeat (SSR) loci were detected, with mononucleotide repeats constituting the majority (63.56%) and showing a pronounced A/T bias (97.45%). We identified 40 long repeat sequences, comprising 20 forward repeats and 20 palindromic repeats. Codon usage bias analysis showed that the average G+C content at the third codon position (GC_3, 27.59%) was significantly lower than that at the first (GC_1, 46.85%) and second (GC_2, 39.50%) positions. The mean value of *ENC* was 44.57, indicating that the codon bias was relatively weak. The results of the neutrality regression analysis showed that the adjusted coefficient of determination (R^2_{adj}) value was 0.016 0, indicating that natural selection was the dominant evolutionary force (with a contribution rate of 91.47%). Ninety percent of optimal codons ended with A/T, consistent with the bias pattern identified by parity rule 2 (PR2) analysis: G content at the third position (G_3) > C content at the third position (C_3), and T content at the third position (T_3) > A content at the third position (A_3). Phylogenetic analysis based on 31 conserved protein-coding genes demonstrated that *Camellia sinensis* cv. Baihaozao formed a highly supported clade (*bootstrap* value = 100%) with *C. sinensis* cv. Fuding Baihao, belonging to the core *Camellia* clade. This study provides references for evaluating tea genetic resources and investigating the evolutionary mechanisms of chloroplast genomes.

Key words: *Camellia sinensis*; chloroplast genome; codon usage bias; optimal codons

茶科(Theaceae)植物是重要的经济作物,其中茶树(*Camellia sinensis*)因其所产茶叶的饮用价值及茶籽的工业用途,在全球范围内具有较好的生态和经济意义[1]。中国作为茶树的起源中心,拥有较高的茶树种质资源多样性,但是目前关于茶树的分类、起源与进化机制仍存在诸多争议[2]。近年来,叶绿体基因组因其结构稳定、进化速率适中且单亲遗传的特性,成为解析植物系统发育、基因表达调控及适应性进化的重要媒介[1-2]。

叶绿体基因组的密码子使用偏好性(Codon usage bias, CUB)是遗传信息传递中的关键特征,反映了同义密码子的非随机使用模式[3]。研究发现,高等植物叶绿体基因组普遍存在第 3 密码子位点的腺嘌呤(A)/胸腺嘧啶(T)使用偏好性,这可能与突变压力和自然选择的共同作用有关[1,4]。例如,在茶科植物中,绝大多数(约 96.55%)高频使用的密码子(即同义密码子中明显更受偏爱的类型)以碱基 A 或 T 结尾。这种密码子的偏好模式越强,基因的表达水平就越高[1]。然而,目前对特定茶树品种(如茶树品种白毫早)叶绿体基因组 CUB 的研究仍较匮乏,其基因组结构特征与密码子偏好模式的关联尚未明晰[2,5]。

高通量测序分析结果表明,不同茶树品种的叶绿体基因组具有高度保守的四分体环状结构,包括大单拷贝区(LSC)、小单拷贝区(SSC)、反向重复区(IRa/IRb)及130~146 个功能基因[含蛋白质编码基因、转运核糖核酸(tRNA)基因及核糖体核糖核酸(rRNA)基因][6]。基因组中同时存在高变异区域(如 *psbA*、*matK*、*petN-psbM*、*trnK-rps16* 等),其序列差异度>1.5%,可作为系统发育分子标记用于物种分类与遗传分化研究[4]。然而,目前关于重要的茶树栽培品种白毫早叶绿体基因组的组装、注释及密码子偏好性的分析仍未见系统报道。此外,现有研究

多聚焦于物种间的比较,关于品种内 CUB 的驱动因素(如自然选择与突变压力的相对贡献)及其对基因异源表达的指导意义仍需深入探讨[7-8]。

本研究拟以茶树品种白毫早为研究对象,结合 Illumina 测序与生物信息学工具,系统解析其叶绿体基因组的结构特征、高变异区域及密码子使用模式[7,9]。通过中性绘图、有效密码子数(*ENC*)计算、所有密码子第 3 位碱基中 G+C 的平均含量占第 3 位碱基总含量的比例(GC_{3s}值)分析及奇偶规则 2(PR2)偏倚性检验,明确 CUB 的主要驱动机制。研究结果以期能够丰富茶树的叶绿体基因组数据库,为茶树的分子育种、系统进化分析及功能基因挖掘奠定基础。

1 材料与方法

1.1 试验材料

茶树品种白毫早(*Camellia sinensis* cv. Baihaozao)的幼嫩叶片样本于 2022 年 6 月采自湖南省娄底市新化县川坳村桃花源农业开发有限公司的标准化种植园(111.28°E,27.73°N)。样本经液氮速冻后用干冰保存,并转运至南京集思慧远生物科技有限公司进行叶绿体基因组测序。本研究所获得的叶绿体基因组序列已提交至美国国家生物技术信息中心(NCBI)数据库(登录号:PQ066596),数据开放获取。

1.2 试验方法

1.2.1 DNA 提取与测序 采用改良十六烷基三甲基溴化铵(CTAB)法提取茶树叶片总脱氧核糖核酸(DNA)[10-11],经纯度及完整性检测合格后,用超声波机械打断法进行 DNA 的片段化处理。片段化产物经末端修复、3′端加 A、测序接头连接及琼脂糖凝胶电泳筛选后,构建 Illumina 测序文库。文库质检合格后,用 Illumina NovaSeq 6000 平台进行双端测序。原始数据用 fastp(v 0.20.0)进行质量控制,具体步骤如下:(1)剔除含接头及引物序列的短序列数据片段(Reads);(2)移除整体碱基平均错误率高于 31.6%的低质量测序数据;(3)去除 N 端碱基数超过 5 个的低置信度 Reads[10]。获得的高质量测序数据用于后续组装与分析。

1.2.2 叶绿体基因组组装 为了降低序列的复杂性,首先用 Bowtie2 v2.2.4 的高灵敏度局部比对模式(Very-sensitive-local)对原始测序数据与自建叶绿体基因组数据库进行比对,筛选出特异性叶绿体 DNA 序列[12]。然后,基于二代测序数据特征,采用无参考基因组的从头组装(De novo)策略,用 SPAdes v3.10.1 软件将 DNA 序列分割成不同 *k-mer* 值(55、87、121)的连续短片段进行迭代组装,获得初始输入(SEED)序列[13]。针对由基因组重复序列造成的组装间隙,用 SSPACE v2.0 构建支架(Scaffolds),并用 GapFiller v2.1.1进行间隙填补[14]。对于持续性缺口区域,采用引物设计软件(Primer 5.0)结合聚合酶链式反应(PCR)扩增技术验证的方法进行序列补全,确保基因组的完整性。最后,用 Bowtie2 对组装序列进行二次比对校正,并基于叶绿体典型的环状四分体结构[大单拷贝区(LSC)、小单拷贝区(SSC)及反向重复区(IRa/IRb)]对基因组坐标进行标准化重排,以消除因反向重复区边界滑动导致的基因位置偏移[15]。

1.2.3 叶绿体基因组的注释 本研究基于 Geneious 平台,采用双策略整合分析流程对叶绿体基因组进行系统注释。首先,运用 prodigal v2.6.3 预测蛋白质的编码区,通过 hmmer v3.1b2 鉴定核糖体 RNA(rRNA)编码区,并用 aragorn v1.2.38 预测转运 RNA(tRNA)编码区[12]。同时,基于 NCBI 数据库获取近缘物种的叶绿体基因序列,用 BLAST v2.6 进行同源比对,获取保守基因的注释结果[12]。然后,通过人工校验消除注释差异,重点修正起始密码子/终止密码子定位误差、反密码子错误及基因冗余问题[16],并用 tRNAscan-SE 1.23 对 tRNA 进行二次验证[11,17]。针对反向重复序列(IR)区域,采用双重同源性阈值法验证其边界完整性,确保发生转剪接的基因座(或产生转剪接转录本的基因)及 rRNA 基因的定位正确[16]。最后,整合多证据注释结果,精确定义多外显子基因的剪接位点,形成标准化基因组注释文件[12,16],并借助 IRscope 工具实现基因组结构的可视化分析。

1.2.4 叶绿体基因组的特征分析 散在重复序列(Dispersed repeats)用 vmatch v2.3.0 软件结合开源动态脚本编辑语言(Perl)进行系统鉴定,参数设置如下:最小重复单元长度为 30 bp,海明距离阈值为 3,涵盖正向、回文、反向及互补 4 种重复类型[16-17]。简单重复序列(SSR)的分析用 MISA v1.0 工具完成,长重复序列的分析借助 REPuter,检测标准为单核苷酸重复次数≥8 次,二核苷酸重复次数≥5 次,三核苷酸至六核苷酸重复次数均≥3 次,以明确叶绿体简单重复序列的分布特征[11,16]。基于近缘物种叶绿体基因组数据,用 CGVIEW 软件进行全基因

组结构的共线性比较,以揭示基因组重排及重复序列的演化规律[17]。

1.2.5 密码子偏好性分析

1.2.5.1 中性分析 使用密码子第 3 位碱基中 G+C 的平均含量(GC_3值)与密码子第 1、第 2 位碱基中 G+C 的平均含量(GC_{12}值)之间的线性回归模型,评估突变压力和自然选择在密码子使用偏好中的相对贡献。如果回归斜率趋近 1 且决定系数(R^2)>0.5,表明密码子偏好性由突变压力主导;反之则提示自然选择占主导地位[18]。

1.2.5.2 *ENC*-GC_3分析 基于 *ENC* 值与 GC_3值的分布关系,通过理论曲线 $ENC = 2 + GC_3 + 29/[GC_3^2 + (1-GC_3)^2]$评估核苷酸组成对密码子使用的限制效应。若数据点位置显著低于理论曲线,表明存在翻译优化或自然选择压力[19]。

1.2.5.3 PR2 分析 以密码子第 3 位碱基中 G 含量占 G+C 含量的比例[$G_3/(G_3+C_3)$]和密码子第 3 位碱基中 A 含量占 A+T 含量的比例[$A_3/(A_3+T_3)$]的偏离程度为指标,基于二等分规则(PR2)的偏差图,评估突变压力与选择压力在功能位点间的非对称程度。中心点(0.5,0.5)用于表征无偏状态,显著偏离中心点则反映选择或突变压力具有非对称作用[18, 20]。

1.2.5.4 最优密码子的筛选 依据 *ENC* 值降序排列,取前十分位组(*ENC* 值≥90%分位数)作为高表达组,后十分位组(*ENC* 值≤10%分位数)作为低表达组,用 CodonW 计算相对同义密码子使用度的变化(△*RSCU*),$\triangle RSCU = RSCU_{高} - RSCU_{低}$。最优密码子的筛选标准为△*RSCU*≥0.08 且$RSCU_{高}$>1.00,以确定受翻译优化驱动的核心最优密码子[21-22]。

1.2.6 叶绿体基因组的系统发育分析 首先对环形叶绿体基因组进行同源起点校正,用 MAFFT 软件(v7.427)的自动优化模式进行多序列比对,该算法利用动态选择比对策略提升序列的匹配精度[9]。随后基于广义时间可逆 Gamma 模型(GTRGAMMA 核苷酸替代模型),用 RAxML v8.2.10 软件进行最大似然法分析,设置快速自举检验(Rapid bootstrap)重复次数为1 000次,以评估节点支持率,构建系统发育树[6,23]。

2 结果与分析

2.1 茶树品种白毫早叶绿体基因组的结构特征

本研究基于 Illumina NovaSeq 6000 高通量测序平台对白毫早茶树(*Camellia sinensis* cv. Baihaozao)的叶绿体基因组进行深度测序。结果显示,测序数据质量优异,质量数大于 20 的碱基所占百分比(Q_{20})和质量数大于 30 的碱基所占百分比(Q_{30})分别达到 97.00%、92.27%(对应碱基错误率分别低于 1.0%、0.1%),累计获得大小为16 140 581 400 bp 的高质量序列,为后续基因组组装提供了基础[24]。

白毫早茶树的叶绿体基因组呈现典型的四分体环状结构,总长度为157 025 bp,由大小为86 586 bp 的大单拷贝区(LSC)、大小为 18 277 bp 的小单拷贝区(SSC)和 1 对大小各为26 081 bp 的反向重复区(IRa/IRb)组成。其全基因组的 G+C 含量为 37.30%,与菊科植物藜蒿(*Artemisia selengensis*)物种具有较高的相似度(相似度为 37.5%)[25]。区域 G+C 含量分析结果显示,IR 区富含核糖体 RNA 基因,G+C 含量最高(42.95%),显著高于 LSC 区(35.33%)、SSC 区(30.55%),该特征与木姜子属(*Litsea*)植物(IR 区的G+C 含量为 42.68%)、腺果藤科(Biebersteiniaceae)植物(IR 区的 G+C 含量为 42.68%)的叶绿体基因组规律一致[26-27]。值得注意的是,SSC 区的 G+C 含量(30.55%)接近清风藤属(*Sabia*)物种的 G+C 含量(32.00%),表明该区域在进化过程中可能承受了更强的 A/T 碱基选择压力[28]。

2.2 茶树品种白毫早叶绿体基因组的基因组成

白毫早茶树叶绿体基因组的基因组成分析结果显示,其环状基因组共注释了 133 个功能基因,包括 87 个蛋白质编码基因、37 个转运核糖核酸(tRNA)基因、8 个核糖体核糖核酸(rRNA)基因及 1 个假基因(表 1),基因总数及功能分类(包括蛋白质编码基因、tRNA 基因、rRNA 基因和假基因)在系统发育近缘物种间高度保守,体现了叶绿体基因组的保守结构特征[29-31]。由表 1 可知,光系统Ⅰ亚基与光系统Ⅱ亚基构成光能捕获与转化的核心单元。其中光系统Ⅰ亚基编码基因包含 *psaA*、*psaB* 等 5 个,参与光能吸收与电子传递;光系统Ⅱ亚基编码基因则包含 15 个,负责水的光解与氧气释放。值得注意的是,光系统Ⅱ亚基编码基因数量明显多于光系统Ⅰ亚基编码基因,暗示其在光反应中承担更复杂的调控功能。在能量代谢相关基因中,NADH 脱氢酶亚基编码基因(*ndh* 基因家族)包含 12 个成员,其中 *ndhB* 有 2 个拷贝,*ndhA*、*ndhB* 各有 1 个内含子,提示其可能通过基因重复与可变剪接实现功能的多样性,该基因可能在电子传递链中起到关键作用。细胞色素 b/f 复合体编码基因(*pet* 基

因家族)与 ATP 合酶亚基编码基因(*atp* 基因家族)各包含 6 个,共同构成质子梯度驱动的 ATP 合成系统。

白毫早茶树叶绿体 *clpP*、*ycf3* 基因各含有 2 个内含子,*atpF*、*ndhA*、*ndhB* 等基因含有单内含子。白毫早茶树叶绿体基因组的内含子长度差异显著,其中 *trnK-UUU* 的内含子最大,大小为2 488 bp,*trnL-UAA* 的内含子最小,大小为 523 bp;外显子长度差异亦明显,其中 *trnG-GCC* 的外显子长度仅为 60 bp,而 *ycf2* 的外显子长度达6 897 bp,提示不同基因在进化过程中受到的选择压力不同[32-33]。

此外,LSC 区的基因密度(143 个基因中含有 35 个外显子、19 个内含子)显著高于 SSC 区和 IR 区,与多数被子植物叶绿体基因组的区域功能分工模式相符[34]。假基因的存在及部分基因内含子数量的变异可能反映基因组重排或自然选择驱动的功能存在冗余[31]。本结果可为后续密码子偏好性分析及系统发育研究提供结构基础,亦可为茶树属叶绿体基因组演化研究补充数据[6]。

表 1 茶树品种白毫早叶绿体基因组的组成

Table 1 Gene composition of the chloroplast genome of *Camellia sinensis* cv. Baihaozao

基因编码产物/功能	基因名称	数量(个)
光系统Ⅰ亚基	*psaA*、*psaB*、*psaC*、*psaI*、*psaJ*	5
光系统Ⅱ亚基	*psbA*、*psbB*、*psbC*、*psbD*、*psbE*、*psbF*、*psbH*、*psbI*、*psbJ*、*psbK*、*psbL*、*psbM*、*psbN*、*psbT*、*psbZ*	15
NADH 脱氢酶亚基	*ndhA**、*ndhB**(2)、*ndhC*、*ndhD*、*ndhE*、*ndhF*、*ndhG*、*ndhH*、*ndhI*、*ndhJ*、*ndhK*	12
细胞色素 b/f 复合体亚基	*petA*、*petB**、*petD**、*petG*、*petL*、*petN*	6
ATP 合酶亚基	*atpA*、*atpB*、*atpE*、*atpF**、*atpH*、*atpI*	6
核糖双加氧酶大亚基	*rbcL*	1
大核糖体亚基蛋白	*rpl14*、*rpl16**、*rpl2**(2)、*rpl20*、*rpl22*、*rpl23*(2)、*rpl32*、*rpl33*、*rpl36*	11
小核糖体亚基蛋白	*rps11*、*rps12***(2)、*rps14*、*rps15*、*rps16**、*rps18*、*rps19*、*rps2*、*rps3*、*rps4*、*rps7*(2)、*rps8*	14
RNA 聚合酶亚基	*rpoA*、*rpoB*、*rpoC1**、*rpoC2*	4
核糖体 RNA	*rrn16*(2)、*rrn23*(2)、*rrn4.5*(2)、*rrn5*(2)	8
转运 RNA	*trnA-UGC**(2)、*trnC-GCA*、*trnD-GUC*、*trnE-UUC*、*trnF-GAA*、*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*、*trnfM-CAU*	37
成熟酶	*matK*	1
蛋白酶	*clpP***	1
包膜膜蛋白	*cemA*	1
乙酰辅酶 A 羧化酶	*accD*	1
c 型细胞色素合成基因	*ccsA*	1
翻译起始因子	*infA*	1
保守的假设叶绿体开放阅读框	*ycf1*#、*orf42*(2)、*ycf1*、*ycf2*(2)、*ycf3***、*ycf4*	8

基因名称后面标有 * 表示该基因有 1 个内含子,标有 ** 表示该基因有 2 个内含子,标有(2)表示该基因有 2 个拷贝,标有#表示该基因为假基因。NADH 表示还原型烟酰胺腺嘌呤二核苷酸,ATP 表示三磷酸腺苷,RNA 表示核糖核酸。

2.3 简单重复序列(SSR)和长重复序列

本研究对白毫早茶树叶绿体基因组中的简单重复序列(SSR)和长重复序列进行了系统分析。由表 2 可以看出,从叶绿体基因组中共检测到 247 个 SSR 位点,其中单核苷酸重复(63.56%)占绝对优势,显著高于二核苷酸重复(1.62%)、三核苷酸重复(29.15%)、四核苷酸重复(4.86%)及五核苷酸重复(0.81%)。值得注意的是,单核苷酸重复中 A/T 型占比高达 97.45%,这种 A/T 偏倚性与叶绿体基因组整体表现出的较高的 A、T 含量特征[35-36]一致,反映了植物叶绿体 SSR 的进化保守性[37]。

由表 3 可以看出,通过对长重复序列分析,共鉴定出 40 个长重复序列,包括 20 个正向重复序列和 20 个回文重复序列,未检测到反向重复序列或互补重复序列。其中,正向重复序列的长度为30~82 bp。这一分布模式与部分物种中同时存在正向、反向及回文重复序列的研究结果[38]形成对比,提示白毫早茶树可能存在叶绿体基因组长重复序

列类型的特异性。此外,回文重复的序列长度跨度较大,范围为30~26 081 bp,可能与其在基因组结构重排或调控元件中的功能相关[37]。

表 2 茶树品种白毫早叶绿体基因组中简单重复序列(SSR)的类型及数量分布

Table 2 Types and quantity distribution of simple sequence repeats (SSRs) in the chloroplast genome of *Camellia sinensis* cv. Baihaozao

重复单元碱基类型	重复次数															总 SSR 数量(个)
	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	
A/T						58	44	14	15	9	6	4	1	1	1	153
C/G						4										4
AT/AT			4													4
AAT/AAG/AAC/ACC/AGA/AGC/ATA/ATC/ATG/ATT	25															25
TAA/TCA/TCT/TGA/TGC/TTA/TTC/TTG	27	1														28
CAA/CAG/CTC/CTG/CTT	9															9
GAA/GAG/GAT/GCA/GCT/GGA/GGT	10															10
AAAT/AATA/AGAT/ATAG	5															5
CCCT	1															1
GAAA/GAGG/GTCT	3															3
TCTA/TCTT/TTTC	3															3
AAAAAG/CTTTTT	2															2

表 3 茶树品种白毫早叶绿体基因组中长重复序列

Table 3 Long repeat sequences in the chloroplast genome of *Camellia sinensis* cv. Baihaozao

编号	第 1 部分的重复长度(bp)	第 1 部分的起始位置(bp)	第 2 部分的重复长度(bp)	第 2 部分的起始位置(bp)	匹配方向	2 个重复序列间的距离(bp)	*E* 值
1	26 081	86 587	26 081	130 945	P	0	0
2	82	93 854	82	93 872	F	−3	7.09×10^{-34}
3	82	93 854	82	149 659	P	−3	7.09×10^{-34}
4	82	93 872	82	149 677	P	−3	7.09×10^{-34}
5	82	149 659	82	149 677	F	−3	7.09×10^{-34}
6	70	93 866	70	93 884	F	−2	1.08×10^{-28}
7	70	93 866	70	149 659	P	−2	1.08×10^{-28}
8	70	93 884	70	149 677	P	−2	1.08×10^{-28}
9	60	93 876	60	93 894	F	−1	9.39×10^{-25}
10	60	93 876	60	149 659	P	−1	9.39×10^{-25}
11	60	93 894	60	149 677	P	−1	9.39×10^{-25}
12	60	93 854	60	93 890	F	−3	4.82×10^{-21}
13	60	93 854	60	149 663	P	−3	4.82×10^{-21}
14	60	93 890	60	149 699	P	−3	4.82×10^{-21}
15	60	149 659	60	149 695	F	−3	4.82×10^{-21}
16	52	93 866	52	93 902	F	−2	4.08×10^{-18}
17	50	76 695	50	76 695	P	−2	6.03×10^{-17}
18	42	79 146	42	79 146	P	0	3.59×10^{-16}
19	42	100 937	42	122 869	F	0	3.59×10^{-16}
20	42	122 869	42	142 634	P	0	3.59×10^{-16}

续表3 Continued3

编号	第 1 部分的重复长度(bp)	第 1 部分的起始位置(bp)	第 2 部分的重复长度(bp)	第 2 部分的起始位置(bp)	匹配方向	2 个重复序列间的距离(bp)	*E* 值
21	42	93 876	42	93 912	F	−1	4.52×10^{-14}
22	42	45 582	42	122 868	F	−3	1.11×10^{-10}
23	42	93 854	42	93 908	F	−3	1.11×10^{-10}
24	42	149 659	42	149 713	F	−3	1.11×10^{-10}
25	39	45 585	39	100 939	F	−2	1.53×10^{-10}
26	39	45 585	39	142 635	P	−2	1.53×10^{-10}
27	38	60 758	38	60 774	F	−3	2.09×10^{-8}
28	35	40 539	35	42 763	F	−3	1.04×10^{-6}
29	34	93 866	34	93 920	F	−2	1.19×10^{-7}
30	32	9 048	32	37 381	F	−3	5.03×10^{-5}
31	31	38 704	31	38 704	P	−3	1.83×10^{-4}
32	30	9 050	30	47 298	P	0	6.01×10^{-9}
33	30	14 203	30	14 203	P	−2	2.35×10^{-5}
34	30	37 383	30	47 298	P	−3	6.59×10^{-4}
35	30	45 597	30	100 951	F	−3	6.59×10^{-4}
36	30	45 597	30	142 632	P	−3	6.59×10^{-4}
37	30	91 418	30	91 460	F	−3	6.59×10^{-4}
38	30	91 418	30	152 123	P	−3	6.59×10^{-4}
39	30	91 460	30	152 165	P	−3	6.59×10^{-4}
40	30	152 123	30	152 165	F	−3	6.59×10^{-4}

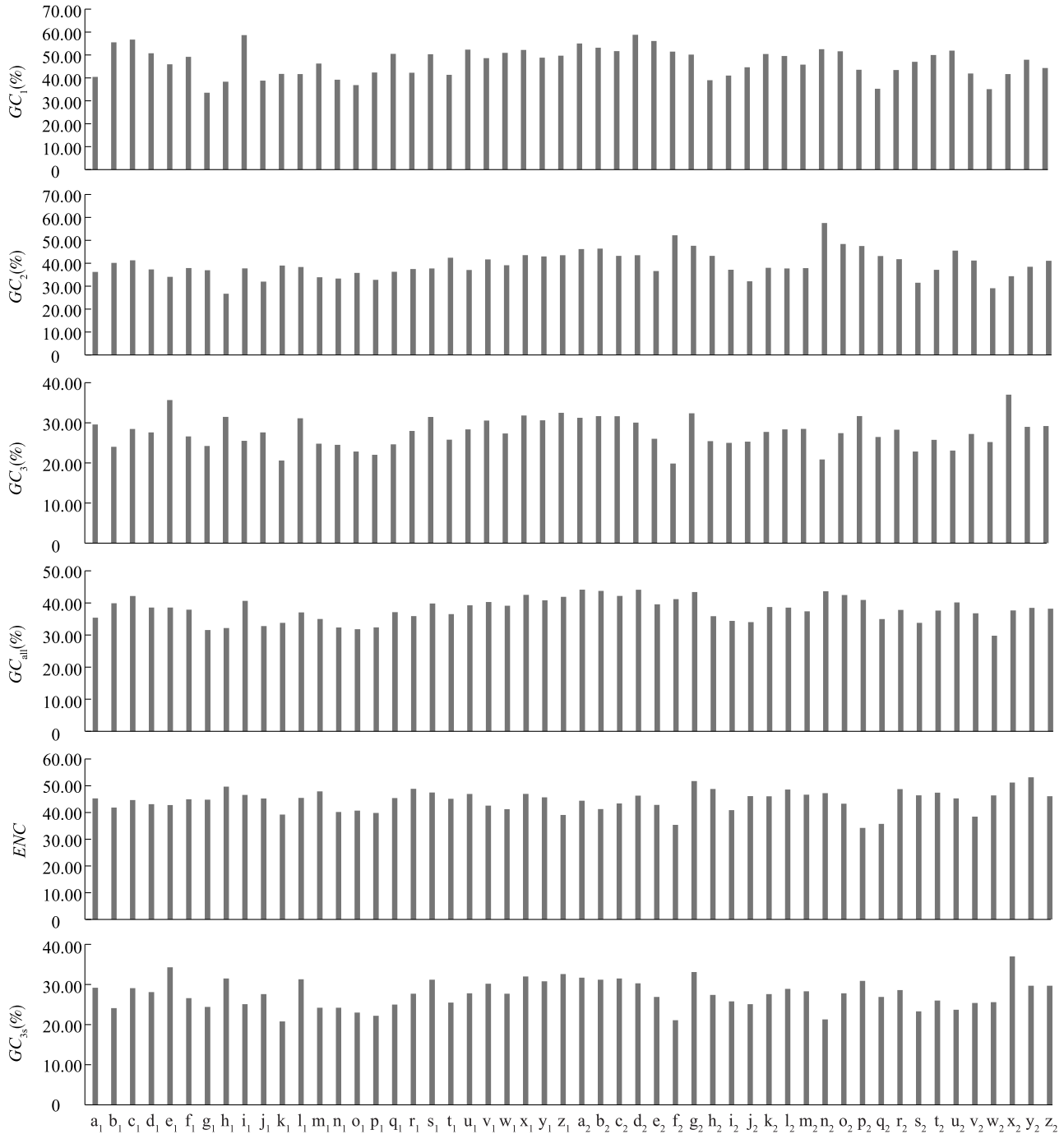
P:回文重复; F:正向重复;*E* 值:评估 2 个序列比对结果统计学显著性的指标,当 *E* 值小于 1.0×10^{-10}时,认为比对结果具有显著性意义。

与已报道的植物叶绿体 SSR 特征相比,本研究发现的 SSR 总数(247 个)显著高于豆科胡枝子属(*Lespedeza*)植物(75~78 个单核苷酸重复)[11],但低于甘蔗(2 199个单核苷酸重复)[39],表明 SSR 的丰度存在显著的物种特异性。值得注意的是,四核苷酸重复(12 个)和五核苷酸重复(2 个)的检出,补充了部分研究中“叶绿体 SSR 以单核苷酸/双核苷酸/三核苷酸为主”的结果[40-41]。

2.4 茶树品种白毫早叶绿体基因组密码子使用偏好性

2.4.1 同义密码子的偏好性 由图 1 可以看出,白毫早茶树叶绿体基因组中 52 个蛋白质编码基因的 GC_3均值为 27.59%,显著低于 GC_1的均值(46.85%)和 GC_2的均值(39.50%),呈现 $GC_1>GC_2>GC_3$的规律。该模式与多种被子植物[如向日葵(*Helianthus annuus* J-01)[42]]及单子叶植物叶绿体基因组的碱基组成特征[43-44]一致,反映了第 3 位密码子对 A/T 的显著偏好,可能与叶绿体基因组中普遍存在的 A/T 偏向性突变压力相关。进一步分析发现,基因组的 *ENC* 值为 34.24~53.16,均值为 44.57,其中 51 个基因的 *ENC* 值>35.00,表明整体密码子的偏好性较弱,提示同义密码子的使用主要受到突变-漂变平衡影响而非受到强选择性压力影响[42,45]。值得注意的是,仅 *rps14* 基因的 *ENC* 值(34.24)显著低于其他基因,可能由于其功能需求(如翻译效率或 mRNA 稳定性)受到局部自然选择的影响。

2.4.2 中性分析(GC_3-GC_{12}分析) 如表 4 所示,白毫早茶树叶绿体基因组 GC_3值(19.85%~37.02%)整体低于 GC_{12}值(32.11%~55.04%),该差异与茶科植物叶绿体基因组中普遍观察到的 GC_3偏向性一致,回归分析结果显示其差异主要源于弱选择约束下的 A/T 偏向突变压力[1,46]。通过构建 $GC_{12}(Y)$与 $GC_3(x)$的中性回归模型,发现二者的线性回归方程为 $Y=0.408\ 0+0.085\ 3x$,对应的校正决定系数(R^2_{adj})= 0.016 0,回归斜率为0.085 3,表明自然选择对密码子偏好形成的贡献率为 91.47%,显著高于突变压力的贡献(8.53%)[47-48]。低 R^2_{adj}($R^2_{adj}<0.050\ 0$)提示 GC_{12}与 GC_3间无显著相关性,进一步支持叶绿体基因组密码子的偏好性主要受到自然选择驱动而非受到突变偏置主导的假设[44,47]。该研究结果反映了叶绿体基因在进化过程中受翻译效率优化等选择压力约束的保守性[1,49]。



a_1:*accD*;b_1:*atpA*;c_1:*atpB*;d_1:*atpE*;e_1:*atpF*;f_1:*atpI*;g_1:*ccsA*;h_1:*cemA*;i_1:*clpP*;j_1:*matK*;k_1:*ndhA*;l_1:*ndhB*;m_1:*ndhC*;n_1:*ndhE*;o_1:*ndhF*;p_1:*ndhG*;q_1:*ndhH*;r_1:*ndhI*;s_1:*ndhJ*;t_1:*ndhK*;u_1:*petA*;v_1:*petB*;w_1:*petD*;x_1:*psaA*;y_1:*psaB*;z_1:*psbA*;a_2:*psbB*;b_2:*psbC*;c_2:*psbD*;d_2:*rbcL*;e_2:*rpl14*;f_2:*rpl16*;g_2:*rpl2*;h_2:*rpl20*;i_2:*rpl22*;j_2:*rpoA*;k_2:*rpoB*;l_2:*rpoC1*;m_2:*rpoC2*;n_2:*rps11*;o_2:*rps12*;p_2:*rps14*;q_2:*rps18*;r_2:*rps2*;s_2:*rps3*;t_2:*rps4*;u_2:*rps7*;v_2:*rps8*;w_2:*ycf1*;x_2:*ycf2*;y_2:*ycf3*;z_2:*ycf4*。GC_1:密码子第 1 位碱基中 G+C 的平均含量;GC_2:密码子第 2 位碱基中 G+C 的平均含量;GC_3:密码子第 3 位碱基中 G+C 的平均含量;GC_{all}:密码子所有碱基位中 G+C 的平均含量;*ENC*:有效密码子数;GC_{3s}:所有密码子第 3 位碱基中 G+C 含量占第 3 位碱基总含量的比例。

图 1 G+C 含量与有效密码子数(*ENC* 值)在茶树品种白毫早叶绿体基因间的对比分析结果

Fig.1 Comparative analysis results of G+C content and effective number of codons (*ENC* values) among genes in the chloroplast genome of *Camellia sinensis* cv. Baihaozao

2.4.3 GC_3-*ENC* 分析 由表 4 还可以看出,52 个基因的 *ENC* 值多数分布在 39.00~54.00,平均值为 44.57,其中 29 个基因的 *ENC* 值>45.00,表明白毫早茶树叶绿体基因组的整体密码子偏好性较弱[50]。值得注意的是,所有基因的实测 *ENC* 值均极显著低于基于 GC_3组成的理论预期值[假设当密码子的使用仅由突变压力(即 G+C 含量)驱动,不受自然选择影响时,通过数学模型计算出的 *ENC* 的预测值](*P*<0.01),提示密码子使用模式受到非突变压力因素的强烈影响[51-52]。GC_3 值分布范围狭窄(19.85%~37.02%),这一现象与茶科植物(Theaceae)及向日葵(*Helianthus annuus*)叶绿体基因组的 GC_3分布模式相似,均呈现对 A/T 的偏好性[1]。

表 4 茶树品种白毫早叶绿体基因组密码子分析结果

Table 4 Analysis of codons in the chloroplast genome of *Camellia sinensis* cv. Baihaozao

基因编码产物	基因	GC_{12}	GC_3	*ENC*	$G_3/(G_3+C_3)$	$A_3/(A_3+T_3)$	基因编码产物	基因	GC_{12}	GC_3	*ENC*	$G_3/(G_3+C_3)$	$A_3/(A_3+T_3)$
乙酰辅酶 A 羧化酶	*accD*	38.33	29.58	45.28	0.54	0.34		*psbB*	50.59	31.24	44.42	0.53	0.39
ATP 合酶亚基	*atpA*	47.84	24.02	41.86	0.55	0.48		*psbC*	49.79	31.65	41.28	0.46	0.44
	atpB	49.00	28.46	44.65	0.53	0.46		*psbD*	47.46	31.64	43.38	0.43	0.37
	atpE	44.03	27.61	43.11	0.62	0.44	核糖双加氧酶大亚基	*rbcL*	51.16	30.04	46.29	0.50	0.42
	atpF	40.00	35.68	42.79	0.62	0.50	大核糖体亚基蛋白	*rpl14*	46.35	26.02	42.84	0.53	0.49
	atpI	43.55	26.61	44.95	0.44	0.43		*rpl16*	51.84	19.85	35.39	0.63	0.59
c 型细胞色素	*ccsA*	35.25	24.22	44.81	0.50	0.46		*rpl2*	48.91	32.36	51.74	0.51	0.49
包膜膜蛋白	*cemA*	32.54	31.47	49.68	0.48	0.43		*rpl20*	41.10	25.42	48.80	0.63	0.49
蛋白酶	*clpP*	48.22	25.51	46.57	0.56	0.49		*rpl22*	39.11	25.00	40.88	0.56	0.54
成熟酶	*matK*	35.40	27.60	45.24	0.54	0.43	RNA 聚合酶亚基	*rpoA*	38.39	25.30	46.13	0.56	0.47
NADH 脱氢酶亚基	*ndhA*	40.39	20.60	39.22	0.49	0.49		*rpoB*	44.21	27.73	46.06	0.64	0.49
	ndhB	40.02	31.12	45.46	0.42	0.46		*rpoC1*	43.64	28.36	48.58	0.55	0.48
	ndhC	40.08	24.79	47.89	0.67	0.41		*rpoC2*	41.84	28.48	46.66	0.53	0.48
	ndhE	36.28	24.51	40.20	0.56	0.35	小核糖体亚基蛋白	*rps11*	55.04	20.86	47.22	0.59	0.48
	ndhF	36.32	22.83	40.69	0.60	0.39		*rps12*	50.00	27.42	43.28	0.44	0.53
	ndhG	37.57	22.03	39.83	0.54	0.41		*rps14*	45.54	31.68	34.24	0.72	0.58
	ndhH	43.40	24.62	45.40	0.61	0.49		*rps18*	39.22	26.47	35.74	0.67	0.47
	ndhI	39.88	27.98	48.87	0.49	0.44		*rps2*	42.62	28.27	48.73	0.58	0.45
	ndhJ	44.03	31.45	47.49	0.60	0.40		*rps3*	39.27	22.83	46.43	0.38	0.60
	ndhK	41.87	25.80	45.14	0.55	0.43		*rps4*	43.57	25.74	47.41	0.48	0.51
细胞色素 b/f 复合体亚基	*petA*	44.71	28.35	46.92	0.56	0.43		*rps7*	48.72	23.08	45.27	0.56	0.58
	petB	45.14	30.56	42.57	0.59	0.40		*rps8*	41.55	27.21	38.44	0.49	0.52
	petD	45.03	27.33	41.26	0.55	0.50	保守的假定叶绿体开放阅读框编码产物	*ycf1*	32.11	25.20	46.38	0.54	0.52
光系统Ⅰ亚基	*psaA*	47.87	31.82	46.96	0.49	0.42		*ycf2*	38.02	37.02	51.20	0.54	0.43
	psaB	45.92	30.61	45.67	0.56	0.38		*ycf3*	43.20	28.99	53.16	0.55	0.47
光系统Ⅱ亚基	*psbA*	46.61	32.49	39.09	0.33	0.33		*ycf4*	42.70	29.19	46.09	0.54	0.41

NADH:还原型烟酰胺腺嘌呤二核苷酸;ATP:三磷酸腺苷;*ENC*:有效密码子数;GC_{12}:密码子第 1、2 位碱基中 G+C 的平均含量;GC_3:密码子第 3 位碱基中 G+C 的平均含量;$G_3/(G_3+C_3)$:密码子第 3 位碱基中 G 含量占 G+C 总含量的比例;$A_3/(A_3+T_3)$:密码子第 3 位碱基中 A 含量占 A+T 总含量的比例。

2.4.4 PR2 分析 根据 PR2 偏倚分析结果,白毫早茶树叶绿体基因组中密码子第 3 位碱基使用偏好性呈现显著特异性。通过计算 $G_3/(G_3+C_3)$ 和 $A_3/(A_3+T_3)$发现,基因密码子第 3 位普遍存在 $G_3>C_3$

[$G_3/(G_3+C_3)$均值>0.50]和$T_3>A_3$[$A_3/(A_3+T_3)$均值<0.50]的特点(表4)。值得注意的是,*psbA*基因的$G_3/(G_3+C_3)$值和$A_3/(A_3+T_3)$值均最低(0.33),表明其密码子第3位对C、T的偏好性最强,这与叶绿体*psbA*基因密码子普遍存在以C结尾的偏好性一致[53]。然而,*rps14*基因的$G_3/(G_3+C_3)$值最高(0.72),*rps3*基因的$A_3/(A_3+T_3)$值最高(0.60),提示核糖体蛋白质编码基因可能受到不同于其他功能基因的选择性压力[27,54]。

2.4.5 *最优密码子的确定* 根据筛选标准(*RSCU*>1.00且△*RSCU*≥0.08),本研究一共从白毫早茶树叶绿体基因组中鉴定出20个最优密码子,其中90%的密码子以A或T结尾(如GCA、GCT、CGA等),比例显著高于以G/C结尾的密码子(表5)。这一结果表明,白毫早茶树叶绿体基因组表现出对A/T结尾密码子的强烈偏好,与蕨类植物铁线蕨(*Adiantum capillus-veneris*)(75%以A/T结尾)、理查德水蕨(*Ceratopteris richardii*)(89%以A/T结尾)等物种的叶绿体密码子使用模式高度一致[51]。进一步分析发现,这种偏好性可能与叶绿体基因组固有的高A/T含量和突变偏置密切相关[44,49],同时自然选择压力可能通过优化翻译效率强化了该趋势[44,55]。值得注意的是,本研究筛选的最优密码子中,GCA(编码丙氨酸)、GCT(编码丙氨酸)等高频密码子与柴胡(*Bupleurum falcatum*)叶绿体基因组中高表达密码子存在重叠[56],提示不同物种间在叶绿体基因表达调控机制方面存在一定的保守性。

表5 茶树品种白毫早叶绿体基因的最优密码子

Table 5 Optimal codons in the chloroplast genome of *Camellia sinensis* cv. Baihaozao

氨基酸	密码子	基因组		高表达基因		低表达基因		△*RSCU*
		数量(个)	*RSCU*	数量(个)	*RSCU*	数量(个)	*RSCU*	
丙氨酸(Ala)	GCA *	42	1.15	17	1.15	25	1.02	0.13
	GCT ***	78	1.84	37	2.51	41	1.67	0.84
精氨酸(Arg)	CGA **	72	1.42	26	1.84	46	1.37	0.47
	CGT **	53	1.39	20	1.41	33	0.99	0.42
半胱氨酸(Cys)	TGT ***	36	1.52	6	2.00	30	1.43	0.57
谷酰胺酸(Gln)	CAA **	101	1.53	19	1.73	82	1.43	0.30
谷氨酸(Glu)	GAA *	153	1.53	33	1.53	120	1.34	0.19
甘氨酸(Gly)	GGT ***	70	1.30	31	1.91	39	1.04	0.87
亮氨酸(Leu)	CTA *	61	0.78	13	1.15	48	0.90	0.25
	CTT *	94	1.25	18	1.59	76	1.43	0.16
	TTA **	82	1.98	18	1.59	64	1.20	0.39
	TTG *	94	1.23	18	1.59	76	1.43	0.16
赖氨酸(Lys)	AAA **	158	1.53	31	1.72	127	1.30	0.42
脯氨酸(Pro)	CCT **	58	1.65	15	1.76	43	1.38	0.38
丝氨酸(Ser)	AGT **	69	1.30	17	1.52	52	1.05	0.47
	TCT **	101	1.79	22	1.97	79	1.60	0.37
苏氨酸(Thr)	ACC **	35	0.74	11	1.07	24	0.71	0.36
	ACT ***	64	1.65	20	1.95	44	1.30	0.65
缬氨酸(Val)	GTA ***	55	1.53	18	1.85	37	1.25	0.60
	GTT ***	59	1.48	19	1.95	40	1.36	0.59

RSCU:相对同义密码子使用频率;△*RSCU*:相对同义密码子使用度差异,△*RSCU* = $RSCU_{高} - RSCU_{低}$;* 表示0.08≤△*RSCU*<0.30;** 表示0.30≤△*RSCU*<0.50;*** 表示△*RSCU*≥0.50。

2.5 系统发育分析

为了明确白毫早茶树的进化地位,本研究选取禾本科的玉蜀黍属(*Zea*)、甘蔗属(*Saccharum*)、小麦属(*Triticum*)及稻属(*Oryza*)植物作为外群,基于31个保守蛋白质编码基因的联合比对数据矩阵,采用最大似然法(ML)构建系统发育树。图2结果显

示,白毫早茶树与福鼎白毫茶树(*C. sinensis* cv. Fuding Baihao)形成高度支持的分支[自举值(*bootstrap*) = 100%],并与山茶属中的茶(*C. sinensis*, OL450428.1)聚为一类,证实其分类地位属于山茶属的核心类群[57-58]。值得注意的是,白毫早茶树所在的山茶属(*Camellia*)与多瓣茶属(*Polyspora*)构成姐妹群,而紫茎属(*Stewartia*)则位于较远分支,表明白毫早茶树与多瓣茶属植物的遗传分化程度低于其与紫茎属植物的分化程度。这一结果与叶绿体基因组的结构保守性特征一致,进一步支持叶绿体蛋白质编码基因在解析山茶属近缘物种关系方面的有效性[57,59]。

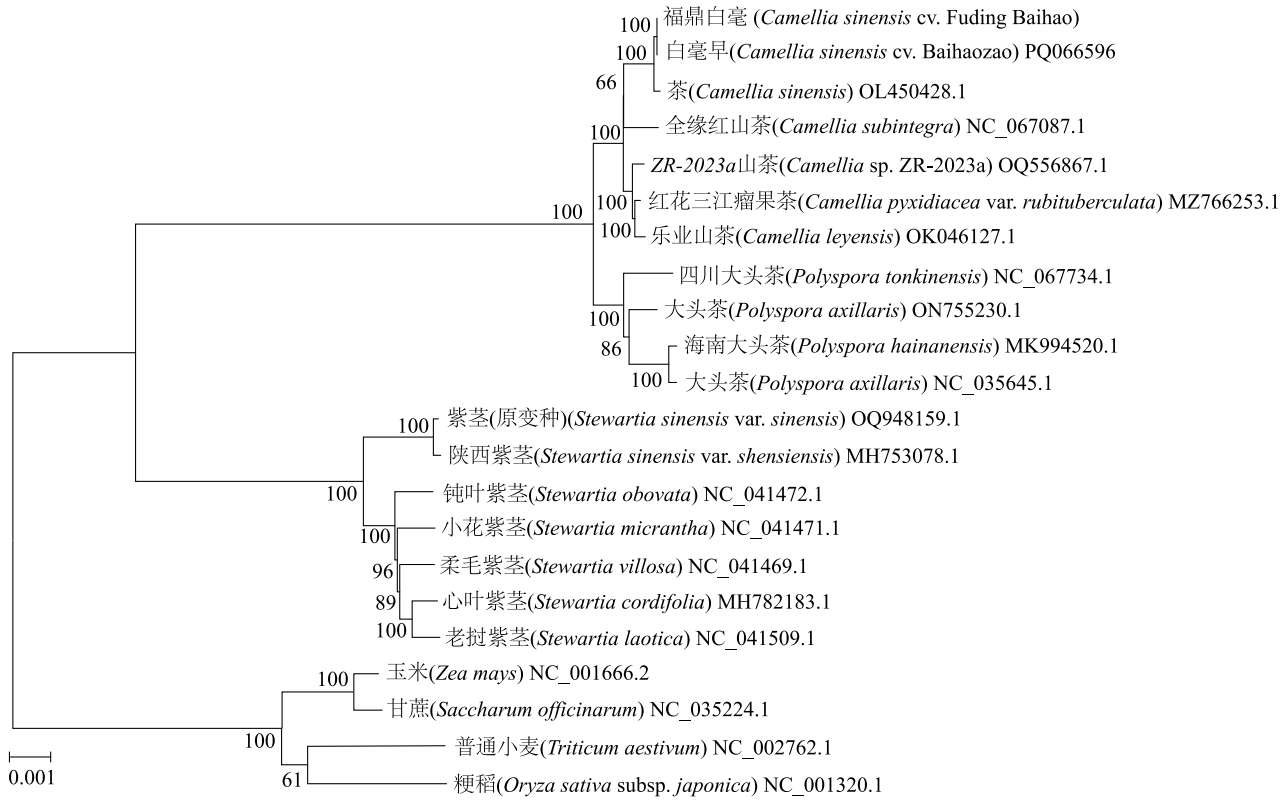


图 2 基于叶绿体基因组的茶树品种白毫早的系统进化分析

Fig.2 Phylogenetic analysis of *Camellia sinensis* cv. Baihaozao based on its chloroplast genome

3 讨 论

本研究揭示了白毫早茶树叶绿体基因组的典型四分体结构(LSC/SSC/IRa/IRb)及其独特的组成特征,其总长度为157 025 bp,与多数被子植物叶绿体基因组规模相当[30,32,60],而 IR 区的 G+C 含量为 42.95%,显著高于 LSC 区(35.33%)和 SSC 区(30.55%)。值得注意的是,基因注释结果显示,*trnK-UUU* 的内含子大小为 2 488 bp,远超多数植物叶绿体基因的内含子长度[如金腰属(*Chrysosplenium*)植物的最大内含子长度仅为 523 bp[61]],可能与其特殊的 RNA 编辑机制相关。

在密码子偏好性方面,GC_3 的均值(27.59%)显著低于 GC_1(46.85%)和 GC_2(39.50%),呈现 $GC_1 > GC_2 > GC_3$ 的递减规律,与双子叶植物线粒体基因组中观察到的嘧啶偏好性(T_3/C_3 优势)[62]一致。*ENC* 值分布(均值为 44.57)及 GC_{12} 与 GC_3 呈低相关性(R^2_{adj} = 0.016 0),表明自然选择的贡献率显著高于突变压力。值得注意的是,最优密码子中 90%以 A/T 结尾,与普通小麦叶绿体基因组中以 A/T 结尾的密码子占优势形成呼应[62],提示翻译效率优化可能驱动密码子选择,这与茶树 *CsNRT1.1* 基因中以 A/T 结尾的密码子占比高达 85%的研究结果相互印证,凸显了山茶科植物叶绿体基因组在密码子选择机制上的趋同演化特征[63]。

SSR 分析结果显示,在 247 个简单重复中,单核

苷酸重复占比为 63.56%(A/T 占绝对优势),与山楂属植物叶绿体基因组的 SSR 类型分布(单核苷酸重复占比为 65.20%)相比更为保守,反映了二者在微卫星序列演化过程中均受到 A、T 碱基优势的强烈驱动[64]。

4 结 论

本研究基于白毫早茶树叶绿体基因组的结构与进化特征研究,为山茶属系统分类提供了分子证据。白毫早茶树叶绿体基因组呈现典型四分体环状结构(总长157 025 bp),包括大单拷贝区(LSC)、小单拷贝区(SSC)和 1 对倒置重复区(IR)。G+C 含量分布具有显著的区域保守性:IR 区(42.95%)因富含 rRNA 基因而较高,SSC 区(30.55%)因 A/T 偏向基因富集而最低。白毫早叶绿体 133 个功能基因的数量与组成与山茶属物种高度一致,印证了该属基因组的进化保守性。SSR 分析结果显示,在单核苷酸重复中单核苷酸 A/T 重复占比 97.45%,此类重复可作为高变异位点用于后续群体遗传研究。密码子偏好性分析结果表明,自然选择是主要驱动因素,贡献率为 91.47%,90%最优密码子以 A/T 结尾,反映叶绿体基因组受到翻译选择压力影响的进化特征。系统发育分析结果进一步证实,白毫早茶树与福鼎白毫茶树形成高度支持的分支(*bootstrap* = 100%),支持其归属于山茶属核心类群。本研究从基因组层面揭示了茶树品种白毫早的分子标记与进化保守性,为山茶属系统分类及叶绿体功能进化研究提供了数据支撑。

参考文献:

[1] WANG Z J, CAI Q W, WANG Y, et al. Comparative analysis of codon bias in the chloroplast genomes of Theaceae species[J]. Frontiers in Genetics,2022,13:824610.

[2] 杨雨青, 谭 娟, 汪 芳, 等. 茶树叶绿体基因组的研究与应用进展[J]. 生物技术通报,2024,40(2):20-30.

[3] 赵 洋,刘 振,杨培迪,等. 密码子偏性分析方法及茶树中密码子偏性研究进展[J]. 茶叶通讯,2016,43(2):3-7.

[4] LI W, ZHANG C P, GUO X, et al. Complete chloroplast genome of *Camellia japonica* genome structures, comparative and phylogenetic analysis[J]. PLoS One,2019,14(5):e0216645.

[5] 徐礼羿. 茶树 SNP 高密度遗传连锁图谱构建与数量性状候选基因挖掘[D]. 武汉:华中农业大学,2019.

[6] CHEN Z Y, LIU Q, XIAO Y, et al. Complete chloroplast genome sequence of *Camellia sinensis*: genome structure, adaptive evolution, and phylogenetic relationships[J]. Journal of Applied Genetics,2023,64(3):419-429.

[7] 王占军,李 豹,姜行舟,等. 两种茶树全基因组数据的密码子偏好性比较分析[J]. 中国细胞生物学学报,2018,40(12):2028-2039.

[8] 王占军,吴子琦,王朝霞,等. 3 个茶树品种 *WOX* 基因家族的进化及密码子偏好性比较[J]. 南京林业大学学报(自然科学版),2022,46(2):71-80.

[9] KATOH K, STANDLEY D M. MAFFT multiple sequence alignment software version 7:improvements in performance and usability[J]. Molecular Biology and Evolution,2013,30(4):772-780.

[10] YU Y F, OUYANG Z, GUO J, et al. Complete chloroplast genome sequence of *Erigeron breviscapus* and characterization of chloroplast regulatory elements[J]. Frontiers in Plant Science, 2021, 12:758290.

[11] SOMARATNE Y, GUAN D L, WANG W Q, et al. The complete chloroplast genomes of two *Lespedeza* species: insights into codon usage bias, RNA editing sites, and phylogenetic relationships in desmodieae (Fabaceae:Papilionoideae)[J]. Plants,2019,9(1):51.

[12] 陆奇丰,黄至欢,骆文华. 极小种群濒危植物广西火桐、丹霞梧桐的叶绿体基因组特征[J]. 生物多样性,2021,29(5):586-595.

[13] GRUENSTAEUDL M, GERSCHLER N, BORSCH T. Bioinformatic workflows for generating complete plastid genome sequences-an example from *Cabomba* (Cabombaceae) in the context of the phylogenomic analysis of the water-lily clade[J]. Life,2018,8(3):25.

[14] 吴东洋. 以杨柳科为例叶绿体基因组组装分析流程管理平台开发及应用[D]. 南京:南京林业大学,2020.

[15] ZHANG T W, ZHANG X W, HU S N, et al. An efficient procedure for plant organellar genome assembly, based on whole genome data from the 454 GS FLX sequencing platform[J]. Plant Methods,2011,7:38.

[16] QU X J, ZOU D, ZHANG R Y, et al. Progress, challenge and prospect of plant plastome annotation[J]. Frontiers in Plant Science,2023,14:1166140.

[17] GICHIRA A W, LI Z Z, SAINA J K, et al. The complete chloroplast genome sequence of an endemic monotypic genus *Hagenia* (Rosaceae):structural comparative analysis, gene content and microsatellite detection[J]. Peer J,2017,5:e2846.

[18] PARVATHY S T, UDAYASURIYAN V, BHADANA V. Codon usage bias[J]. Molecular Biology Reports,2022,49(1):539-565.

[19] CAMIOLO S, MELITO S, PORCEDDU A. New insights into the interplay between codon bias determinants in plants[J]. DNA Research,2015,22(6):461-470.

[20] QIAO Z S, LI J Q, ZHANG X L, et al, Genome-wide identification, expression analysis, and subcellular localization of *DET2* gene family in *Populus yunnanensis*[J]. Genes,2024,15(2):148.

[21] HERSHBERG R, PETROV D A. General rules for optimal codon

choice[J]. PLoS Genetics,2009,5(7):e1000556.

[22] ANGOV E. Codon usage:nature's roadmap to expression and folding of proteins[J]. Biotechnology Journal,2011,6(6):650-659.

[23] STAMATAKIS A. RAxML version 8:a tool for phylogenetic analysis and post-analysis of large phylogenies[J]. Bioinformatics,2014,30(9):1312-1313.

[24] 王兴春,杨致荣,王 敏,等. 高通量测序技术及其应用[J]. 中国生物工程杂志,2012,32(1):109-114.

[25] WANG Y H, WEI Q Y, XUE T Y, et al. Comparative and phylogenetic analysis of the complete chloroplast genomes of 10 *Artemisia selengensis* resources based on high-throughput sequencing[J]. BMC Genomics,2024,25(1):561.

[26] SONG W C, CHEN Z M, SHI W B, et al. Comparative analysis of complete chloroplast genomes of nine species of *Litsea* (Lauraceae):hypervariable regions,positive selection,and phylogenetic relationships[J]. Genes,2022,13(9):1550.

[27] CHI X F, ZHANG F Q, DONG Q, et al. Insights into comparative genomics,codon usage bias,and phylogenetic relationship of species from Biebersteiniaceae and Nitrariaceae based on complete chloroplast genomes[J]. Plants,2020,9(11):1605.

[28] CHEN Q Y, CHEN C L, WANG B, et al. Complete chloroplast genomes of 11 *Sabia* samples:genomic features,comparative analysis,and phylogenetic relationship[J]. Frontiers in Plant Science,2022,13:1052920.

[29] FENG Z, ZHENG Y, JIANG Y, et al. Phylogenetic relationships,selective pressure and molecular markers development of six species in subfamily Polygonoideae based on complete chloroplast genomes[J]. Scientific Reports,2024,14(1):9783.

[30] KIM B, KIM J, PARK H, et al. The complete chloroplast genome sequence of *Bienertia sinuspersici*[J]. Mitochondrial DNA Part B,Resources,2016,1(1):388-389.

[31] CHEN M M, ZHANG M, LIANG Z S, et al. Characterization and comparative analysis of chloroplast genomes in five *Uncaria* species endemic to China[J]. International Journal of Molecular Sciences,2022,23(19):11617.

[32] JIN G Z, LI W J, SONG F, et al. Comparative analysis of complete *Artemisia* subgenus *Seriphidium* (Asteraceae:anthemideae) chloroplast genomes:insights into structural divergence and phylogenetic relationships[J]. BMC Plant Biology,2023,23(1):136.

[33] YAN X K, LIU T J, YUAN X, et al. Chloroplast genomes and comparative analyses among thirteen taxa within Myrsinaceae s.str. clade (Myrsinoideae,Primulaceae)[J]. International Journal of Molecular Sciences,2019,20(18):4534.

[34] ZHAO W, GUO L R, YANG Y, et al. Complete chloroplast genome sequences of *Phlomis fruticosa* and *Phlomoides strigosa* and comparative analysis of the genus *Phlomis* sensu lato (Lamiaceae)[J]. Frontiers in Plant Science,2022,13:1022273.

[35] KAILA T, CHADUVLA P K, RAWAL H C, et al. Chloroplast genome sequence of clusterbean (*Cyamopsis tetragonoloba* L.):genome structure and comparative analysis[J]. Genes,2017,8(9):212.

[36] SHI H W, YANG M, MO C M, et al. Complete chloroplast genomes of two *Siraitia* Merrill species:comparative analysis,positive selection and novel molecular marker development[J]. PLoS One,2019,14(12):e0226865.

[37] RAUBESON L A, PEERY R, CHUMLEY T W, et al. Comparative chloroplast genomics:analyses including new sequences from the angiosperms *Nuphar advena* and *Ranunculus macranthus*[J]. BMC Genomics,2007,8:174.

[38] SOUZA U J B, NUNES R, TARGUETA C P, et al. The complete chloroplast genome of *Stryphnodendron adstringens* (Leguminosae - Caesalpinioideae):comparative analysis with related Mimosoid species[J]. Scientific Reports,2019,9(1):14206.

[39] JIANG H H, WASEEM M, WANG Y, et al. Development of simple sequence repeat markers for sugarcane from data mining of expressed sequence tags[J]. Frontiers in Plant Science,2023,14:1199210.

[40] ASAF S, WAQAS M, KHAN A L, et al. The complete chloroplast genome of wild rice (*Oryza minuta*) and its comparison to related species[J]. Frontiers in Plant Science,2017,8:304.

[41] ASAF S, KHAN A L, KHAN A, et al. Unraveling the chloroplast genomes of two *Prosopis* species to identify its genomic information,comparative analyses and phylogenetic relationship[J]. International Journal of Molecular Sciences,2020,21(9):3280.

[42] CHEN S Y, ZHANG H, WANG X, et al. Analysis of codon usage bias in the chloroplast genome of *Helianthus annuus* J-01[J]. IOP Conference Series:Earth and Environmental Science,2021,792(1):012009.

[43] CAI Z Q, PENAFLOR C, KUEHL J V, et al. Complete plastid genome sequences of *Drimys*,*Liriodendron*,and *Piper*:implications for the phylogenetic relationships of magnoliids[J]. BMC Evolutionary Biology,2006,6:77.

[44] WANG X S, WANG Y Q, LI S H, et al. Analysis of codon usage bias in the *Platycarya* chloroplast genome[J]. Tree Genetics and Molecular Breeding,2021,11(1):1-11.

[45] YANG X, LUO X N, CAI X P. Analysis of codon usage pattern in *Taenia saginata* based on a transcriptome dataset[J]. Parasites & Vectors,2014,7:527.

[46] WANG R, LAN Z, LUO Y J, et al. The complete chloroplast genome of *Stachys geobombycis* and comparative analysis with related *Stachys* species[J]. Scientific Reports,2024,14:8523.

[47] CHEN J, MA W Q, HU X W, et al. Synonymous codon usage bias in the chloroplast genomes of 13 oil-tea *Camellia* samples from South China[J]. Forests,2023,14(4):794.

[48] 赵月梅,徐其碧,杨贵清,等. 艾纳香叶绿体基因组密码子使用偏性分析[J]. 西部林业科学,2023,52(3):55-62,77.

[49] XU C, CAI X N, CHEN Q Z, et al. Factors affecting synonymous codon usage bias in chloroplast genome of *Oncidium* Gower Ramsey[J]. Evolutionary Bioinformatics Online,2011,7:271-278.

[50] 杨洪升,谢 平,李丽丽,等. 杠板归叶绿体基因组密码子偏好

性分析[J/OL]. 分子植物育种[2025-03-05]. https://link.cnki.net/urlid/46.1068.S.20230314.1714.026.

[51] XU P R, ZHANG L J, LU L P, et al. Patterns in genome-wide codon usage bias in representative species of lycophytes and ferns [J]. Genes,2024,15(7):887.

[52] FU Y, LIANG F S, LI C J, et al. Codon usage bias analysis in macronuclear genomes of ciliated protozoa[J]. Microorganisms, 2023,11(7):1833.

[53] 晁岳恩,吴政卿,杨会民,等. 11 种植物 *psbA* 基因的密码子偏好性及聚类分析[J]. 核农学报,2011,25(5):927-932.

[54] PING J, ZHONG X N, WANG T, et al. Structural characterization of *Trivalvaria costata* chloroplast genome and molecular evolution of *rps12* gene in magnoliids[J]. Forests,2023,14:1101.

[55] MORTON B R, SO B G. Codon usage in plastid genes is correlated with context,position within the gene,and amino acid content[J]. Journal of Molecular Evolution,2000,50(2):184-193.

[56] GAO M Q, HUO X W, LU L T, et al. Analysis of codon usage patterns in *Bupleurum falcatum* chloroplast genome[J]. Chinese Herbal Medicines,2023,15(2):284-290.

[57] YANG J B, YANG S X, LI H T, et al. Comparative chloroplast genomes of *Camellia* species[J]. PLoS One,2013,8(8):e73053.

[58] HAO B Q, XIA Y Y, ZHANG Z Y, et al. Comparative analysis of the complete chloroplast genome sequences of four *Camellia* species[J]. Brazilian Journal of Botany,2024,47(1):93-103.

[59] LIN P, YIN H F, WANG K L, et al. Comparative genomic analysis uncovers the chloroplast genome variation and phylogenetic relationships of *Camellia* species[J]. Biomolecules,2022,12(10): 1474.

[60] YE X M, HU D N, GUO Y P, et al. Complete chloroplast genome of *Castanopsis sclerophylla* (lindl.) schott: genome structure and comparative and phylogenetic analysis[J]. PLoS One, 2019, 14 (7):e0212325.

[61] WU Z H, LIAO R, YANG T G, et al. Analysis of six chloroplast genomes provides insight into the evolution of *Chrysosplenium* (Saxifragaceae)[J]. BMC Genomics,2020,21(1):621.

[62] 张文娟. 基于密码子水平的生物信息学分析及进化研究[D]. 上海:复旦大学,2006.

[63] 胡振民,万 青,李 欢,等. 茶树 *CsNRT1.1* 基因密码子使用特性分析[J]. 江苏农业学报,2019,35(4):896-903.

[64] 赵振宁,孙浩田,宋雨茹,等. 山楂属植物叶绿体基因组特征与密码子偏好性分析[J]. 江苏农业学报,2023,39(2):504-517.

(责任编辑:徐 艳)