中国中药杂志
*China Journal of Chinese Materia Medica*
ISSN 1001-5302,CN 11-2272/R

# 《中国中药杂志》网络首发论文

题目：全球物种数量最多的叶绿体基因组数据库的建立及应用进展
作者：陈梓媛，华中一，袁媛
DOI：10.19540/j.cnki.cjcmm.20240909.101
收稿日期：2024-08-19
网络首发日期：2024-09-11
引用格式：陈梓媛，华中一，袁媛．全球物种数量最多的叶绿体基因组数据库的建立及应用进展[J/OL]．中国中药杂志．
https://doi.org/10.19540/j.cnki.cjcmm.20240909.101

# 全球物种数量最多的叶绿体基因组数据库的建立及应用进展

陈梓媛[2]，华中一[2]，袁媛[1]✉

(1. 中国中医科学院 医学实验中心，北京 100700；
2. 中国中医科学院 中药资源中心 道地药材品质保障与资源持续利用全国重点实验室，北京 100700)
*通信作者 *袁媛，研究员，博士生导师，主要从事中药鉴定与分子生药学研究，E-mail：y_yuan0732@163.com

**摘要** 叶绿体基因组是研究植物分类、进化和异源生产次生代谢产物以及蛋白药物的重要工具。随着测序技术的发展和测序成本的下降，叶绿体基因组数据得到了快速积累。但已发布的叶绿体基因组相关数据库存在数据不全面、不完整和管理不足以及叶绿体基因组相关信息不准确、不统一等问题，给叶绿体基因组的开发利用带来更多挑战，亟须建立一个能提供全面且可靠信息的叶绿体基因组数据库。该文对全球物种数量最多的叶绿体基因组综合数据库（Chloroplast Genome Information Resource, CGIR）进行简要介绍。该数据库包含了基因组（genomes）、基因（genes）、微卫星序列（SSRs）、DNA 条形码（barcodes），DNA 特征序列（DSSs）5 个模块，目前已收录了 16 717 个物种的 34 923 条叶绿体基因组。基于数据库模块功能，该文还系统总结了该数据库在植物系统发育分析、物种鉴定和叶绿体基因工程的应用进展。未来将对该叶绿体基因组数据库进行持续更新，以期为叶绿体基因组研究提供坚实可靠的数据基础，并进一步推动中药鉴定、资源保护、种质创新等研究。

**关键词** 叶绿体基因组；数据库；DNA 特征序列；应用进展
中图分类号 文献标志码 DOI: 10.19540/j.cnki.cjcmm.20240909.101

# Establishment and application of chloroplast genome database with the largest number of species in the world

CHEN Zi-yuan[2], HUA Zhong-yi[2], YUAN Yuan[1] ✉

(1. *Experimental Research Center, China Academy of Chinese Medical Sciences, Beijing* 100700*, China*; 2. *State Key Laboratory for Quality Ensurance and Sustainable Use of Dao-di Herbs*，*National Resource Center for Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing* 100700*, China*)

**Abstract** The chloroplast genome is an important tool for studying plant classification, evolution, and the heterologous production of secondary metabolites and protein drugs. With advancements in sequencing technology and reductions in sequencing costs, chloroplast genome data have rapidly accumulated. However, existing chloroplast genome databases suffer from issues such as incomplete data, inadequate management, and inconsistent, inaccurate information, posing significant challenges for the development and utilization of the chloroplast genome. Therefore, it is urgently necessary to establish a database that provides comprehensive and reliable chloroplast genome information. This article provides a brief introduction to the Chloroplast Genome Information Resource (CGIR), the most comprehensive chloroplast genome database globally in terms of species coverage. The database, consisting of five modules, *i.e.*, ① genomes, ②genes, ③ simple sequence repeats (SSRs), ④ DNA barcodes, and ⑤ DNA signature sequences (DSSs), currently includes 34 923 chloroplast genome assemblies from 16 717 species. Based on the functionalities of these modules, the article systematically summarizes the progress in the application of the database in plant phylogenetic analysis, species identification, and chloroplast genetic engineering. The chloroplast genome database will be continuously updated in the future to provide a solid and reliable data foundation for chloroplast genome research, further promoting studies on traditional Chinese medicine（TCM）identification, resource conservation, and germplasm innovation.

**Key words** chloroplast genome; database; DNA signature sequences; application progress

收稿日期：2024-08-19
基金项目：中国中医科学院中药资源中心自主选题研究项目（ZZXT202217）；国家杰出青年科学基金项目（82325049）
作者简介：陈梓媛，博士研究生，主要从事中药鉴定与分子生药学研究，E-mail：c_ziyuan@163.com
网络首发时间：2024-09-11 10:24:07 网络首发地址：https://link.cnki.net/urlid/11.2272.R.20240910.1756.002

叶绿体是绿色植物和藻类细胞中一种半自主的细胞器，拥有相对独立的遗传信息，普遍推断其起源于被古真核细胞吞噬后与其内共生的蓝藻[1]。作为植物细胞中重要的能量转化场所，叶绿体通过光合作用，将光能转化为生物能源[2]。此外，叶绿体还参与了氨基酸、脂肪酸和植物激素和大量次生代谢产物的生物合成。这些次生代谢产物与植物的生长发育息息相关[3]。叶绿体基因组不仅为植物的分类和进化研究提供宝贵的数据资源，也逐渐成为异源生产各类次生代谢产物和蛋白药物的重要平台。

随着测序技术的发展和测序成本的下降，叶绿体基因组研究实现了快速发展。1986 年，首次完成烟草 *Nicotiana tabacum* 的叶绿体全基因组测序[4]，目前超过 60%的已报道叶绿体基因组是在过去 5 年完成测序的，2019—2023 年分别公布了 1 934、2 280、2 252、2 593、1 603 条叶绿体基因组。但目前收录叶绿体基因组的各大数据库存在数据收集不全面、不完整以及数据管理不足等问题，见表 1，如综合数据库中的细胞器基因组数据库通常仅收录了少量叶绿体基因组数据，大多数叶绿体基因组仍然分散在不同的核苷酸数据库中；而部分专业数据库仅包含特定分类物种的叶绿体基因组，如收录种子植物叶绿体基因组的 cpGDB[5]和收录藻类的 OGDA[6]。另一部分则仅限于某些数据类型，如收录叶绿体基因组反向重复区（inverted repeat, IR）的 PCIR[7]和简单重复序列（simple sequence repeat, SSR）的 ChloroMitoSSRDB[8]。随着数据的迅速积累，各大数据库物种分类信息的不准确[9]、基因组术语的不统一[10]等问题也给叶绿体基因组的利用带来了重大挑战。因此，迫切需要建立一个全面收集和管理叶绿体基因组的综合数据库。

表1 叶绿体基因组相关数据库

Table 1 List of chloroplast genome-related databases

| 数据库名称 | 数据库链接 | 数据库类型 | 数据库内容 |
|---|---|---|---|
| NCBI | https://www.ncbi.nlm.nih.gov/ | 综合数据库 | 已公布 13 347 个物种的叶绿体基因组 |
| NGDC[11-12] | https://ngdc.cncb.ac.cn/gwh/ | 综合数据库 | 已公布 530 个物种的 894 条叶绿体基因组 |
| cpGDB[5] | https://www.gndu.ac.in/CpGDB/ | 专业数据库 | 已公布 3 823 个种子植物的叶绿体基因组及其基因序列 |
| OGDA[6] | https://ogda.ytu.edu.cn/ | 专业数据库 | 已公布 920 个藻类的 1 321 条叶绿体基因组及其基因序列 |
| ChloroMitoSSRDB[8] | http://www.mcr.org.in/chloromitossrdb | 专业数据库 | 已公布 370 条叶绿体基因组的 90 774 条 SSR 数据 |
| PCIR[7]* | http://biodb.sdau.edu.cn/pcir/index.php | 专业数据库 | 113 个物种叶绿体基因组的 21 443 条 IR 数据 |

注：*表示数据库已停止访问。

## 1 叶绿体基因组综合数据库 CGIR

中国中医科学院分子生药学创新团队联合中国科学院北京基因组研究所（国家生物信息中心）通过系统地进行叶绿体基因组数据人工审编与分子标记开发，构建了叶绿体基因组综合数据库（Chloroplast Genome Information Resource，CGIR, https://ngdc.cncb.ac.cn/cgir），旨在提供综合全面可靠的叶绿体基因组研究资源。截至 2024 年 9 月 2 日，该数据库已公开了来自 16 717 个物种的 34 923 条叶绿体基因组，数据整合了已发布的叶绿体基因组数据和团队自测的 721 种未发表的叶绿体基因组，为迄今为止物种数量最多的叶绿体基因组数据库。为便于数据的共享应用，CGIR 数据库包含了基因组（genomes）、基因（genes）、微卫星序列（SSRs）、DNA 条形码（barcodes），DNA 特征序列（DSSs）五个模块。

### 1.1 基因组模块

#### 1.1.1 物种分类

CGIR 针对每一条叶绿体基因组的物种分类信息进行人工审编，以消除各数据库和数据提交者造成的数据不一致性。审编主要依据全球生物物种名录（Catalogue of Life, https://www.catalogueoflife.org/）进行，并参考 NCBI Taxonomy 数据库、植物物种清单（The Plant List, http://www.theplantlist.org）、被子植物分类系统（APG Ⅳ）、蕨类植物分类系统（PPG I）及藻库（AlgaeBase, https://www.algaebase.org）。最终确定每一条叶绿体基因组物种名正名，同时也收录其他名称，并标注为异名。在确定物种正名的基础上，CGIR 进一步审校了物种更高层次的分类信息，包括类群（group）、纲（class）、目（order）、科（family）、属（genus）。CGIR 将所有植物的分为 5 种类群：被子植物（angiosperm）、裸子植物（gymnosperm）、蕨类植物（pteridophyte）、

苔藓（bryophyte）和藻类（phycophyta）。并且依据《世界功能性植物物种清单》（World Checklist of UsefulPlant Species）对收录植物的功能进行审编，分为 6 类：环境植物（environmental）、食物植物（food）、饲料植物（forage）、材料植物（material）、药用植物（medicine）和有毒植物（poison）。根据《中国药典》和《全国中草药汇编》中记录的药用植物，CGIR 进一步审编了其药材名、药材英文名和药用部位，其中药材名和药材英文名按原书记录，药用部位被进一步审编。

#### 1.1.2 基因组序列

通过查询文献、数据库搜索，CGIR 收集了叶绿体基因组序列并进行人工审编，排除了错误标注为叶绿体基因组的序列，如 NCBI 数据库收录的序列 MK318534.1 实际为线粒体基因组。此外，对于有多个登录号（Accession ID）的同一条叶绿体基因组，CGIR 进行了归并，并标注了其同义叶绿体基因组登录号。导致该情况一般与 NCBI 将叶绿体基因组指定为参考基因组有关，少部分与用户重复上传有关。此外，CGIR 还收载了团队自测的 721 种未发表的叶绿体基因组。

### 1.2 基因模块

根据叶绿体基因组注释文件，CGIR 对基因名称和注释等信息进行了审编。重点对错误大小写、错误拼写、多余字符、同名基因进行人工审编，如 *AtpA*、*aptA*、*atpA1* 等统一校正为 *atpA*，*lhbA* 和 *psbZ* 等同名基因统一为 *psbZ*。并对注释错误、未知功能基因进行重新注释。

### 1.3 DNA 分子标记模块

针对分子标记开发这一最为常见的应用情景，基于叶绿体基因组序列，CGIR 使用生物信息学方法开发了微卫星序列、DNA 条形码和 DSS 3 种分子标记。同时开发了相应的树型视图，方便用户根据分类层级信息快速寻找目标标记，简化了研究人员开发分子标记的流程。

#### 1.3.1 微卫星序列

微卫星序列又称简单重复序列（SSR）或短串联重复序列（STR）序列，通常由 1~6 bp 的碱基重复单元组成，是真核生物基因组中广泛存在的简单重复序列。SSR 根据其序列可分为 3 类：完全型 SSR（perfect）、不完全型 SSR（imperfect）和复合型 SSR（compound）。基于收录的叶绿体基因组序列，CGIR 使用 MISA 对完全型和复合型 SSR 进行鉴定；使用 IMEx 对不完全型 SSR 进行鉴定。2 个及 2 个以上 SSR 之间的距离不大于 100 bp 的，被认定为复合型 SSR。

#### 1.3.2 DNA 条形码

DNA 条形码是利用一段相对较短而标准的 DNA 序列片段[13, 14, 15]，已报道 *rbcL*、*trnL-trnF*、*trnH-psbA*、*matK*、*ycf5*、*ycf1*、*atpB* 等 19 种植物叶绿体基因组 DNA 条形码。CGIR 使用电子 PCR 获取 DNA 条形码，并使用 BLAST 将每一个基因组的 DNA 条形码区域正向引物与反向引物比对到叶绿体基因组上，根据比对位置，截取引物对中间的核酸序列作为 DNA 条形码。对于 DNA 条形码在叶绿体基因组存在多拷贝的情况，需要进行进一步的人工审编后收录。

#### 1.3.3 DNA 特征序列

为了弥补现有分子标记无法兼具通用性和准确性的局限，团队自主设计了一种新型分子标记，命名为 DNA 特征序列（DNA signature sequences, DSS）[16]。DSS 是指与来源于其他分类单元，即与背景分类单元相比，只出现在某个特定分类单元中的 DNA 序列。CGIR 使用自主开发的 IdenDSS 软件进行 DSS 鉴定，从目标分类单元的叶绿体基因组中任选一条，利用滑窗法生成其所有可能存在的、某个固定长度的 k-mer。将生成的 k-mer 与目标分类单元的其他叶绿体基因组序列进行比对，保留在所有叶绿体基因组序列中都出现的 k-mer。将上一步保留的 k-mer 与背景分类单元的所有叶绿体基因组序列进行比对，保留在所有背景分类单元叶绿体基因组序列中都不出现的 k-mer，作为 DSS。

目前 CGIR 利用所收录的、不少于两个叶绿体基因组序列的物种进行了 DSS 鉴定，即选取其同科的其他物种作为背景分类单元，鉴定单个物种的 DSS，其长度为 40 bp。

## 2 CGIR 的应用

### 2.1 植物系统发育分析

叶绿体基因组具有高度保守的结构和基因组成，为单系遗传[17, 18]，不存在基因重组，分子进化速率适中，且不同区域核苷酸替换速率不同，可适用于不同分类阶元的系统发育分析[19]，利用叶绿体基因组可重建植物目级、科级、属级甚至种及种下水平的系统发育关系[20]。

早期阶段，由于叶绿体全基因组数据的获取面临成本及技术上的巨大挑战，在进行系统发育分析时，往往利用单一或少数几个叶绿体 DNA 序列作为标记。其中，由于叶绿体基因组的蛋白编码区序列变异适中，利用其串联蛋白编码序列进行植物的系统发育分析仍十分常见。LI H T 等[21]基于代表 85%科和所有目的 2 881 个叶绿体基因组，利用其 80 个基因进行被子植物的系统发育重建，并使用 62 块可靠的化石推断被子植物的起源和演化时间。MENG W Q 等[22]对中国特有属玉溪石蒜属 *Shoubiaonia* 属下单种玉溪石蒜 *S. yunnanensis* 进行叶绿体基因组测序与组装，并提取了和石蒜科其余 20 属 34 个物种共有的 79 个蛋白编码序列进行系统发育分析，最终确定了玉溪石蒜属在石蒜科中的系统发育地位。NIU T 等[23]基于禾本科 16 个属 31 个种的 69 个蛋白编码序列构建系统发育树，最终确定羊柴 *Corethrodendron fruticosum* 和同科岩黄芪属 *Hedysarum* 近缘种的亲缘关系。LI T 等[24]利用 22 个木莲属 *Manglietia* 物种和 37 个木兰科其他属物种的 77 个叶绿体基因组蛋白编码序列进行系统发育分析，解析木莲属属内物种的分类关系。

虽然单一或少数几个叶绿体 DNA 序列的使用极大地方便了植物系统发育分析，但这些序列所包含的遗传信息通常不足以提供足够高的分辨率，来区分亲缘关系密切的分类群。完整的叶绿体基因组序列为探究亲缘关系更为密切的分类群之间的系统发育关系和植物的物种起源和演化提供了重要的数据资源。罗晗睿等[25]基于胡颓子科 18 个物种的叶绿体全基因组序列进行了系统发育分析，对胡颓子属 *Elaeagnus*、沙棘属 *Hippophae* 和野牛果属 *Shepherdia* 的系统发育关系、起源和演化进行了推断。QI K 等[26]利用叶绿体全基因组构建了蓼科 49 种植物的系统发育树，确定了西伯利亚蓼 *Knorringia sibirica* 在蓼科中的系统发育位置。SHI W B 等[27]通过对 35 个柑橘属 *Citrus* 物种的叶绿体全基因组系统发育分析，将柑橘属划分为 7 个亚属，为柑橘属的分类提供参考依据。ZENG Q 等[28]基于 50 个桑属 *Morus* 的全叶绿体基因组对桑属物种间的系统发育关系进行了探讨，建议将桑属分为 6 种 2 亚种。

CGIR 对叶绿体基因组的收录不是进行单纯的整合，还对其进行了细致且标准的审编，特别是对错误注释的基因序列进行重新注释和校正。作为当前收录物种数量最多的叶绿体基因组数据库，CGIR 为研究人员提供了大量信息更可靠、更统一的数据。同时开发了相应的具有明确分类信息的树形视图，为研究人员提供更多的植物系统发育分析相关信息，简化数据搜索流程。

### 2.2 物种鉴定

#### 2.2.1 SSR 标记

叶绿体 SSR 具有共显性、高多态性和分布广泛等优点，而被广泛应用于种质资源鉴定和遗传多样性分析等研究[29]。李思巧等[30]基于花椒 *Zanthoxylum bungeanum* 的叶绿体基因组 SSR，筛选出 10 对可用于花椒属 *Zanthoxylum* 种间鉴定的多态性引物，其中 2 对引物可以把花椒属的花椒 *Z. bungeanum*、竹叶花椒 *Z. armatum* 和日本花椒 *Z. piperitum* 区分，1 对引物可以将竹叶花椒种内的‘狗椒’区分。严卓彦等[31]利用风藤和山蒟的叶绿体 SSR 差异位点，筛选出了 4 对特异性引物，根据其毛细管电泳检测结果，可对风藤 *Piper kadsura* 和山蒟 *P. hancei* 进行有效的鉴别。即便在 DNA 高度降解的发酵酒中，叶绿体 SSR 依旧是很有应用前景的品种鉴定工具，如 BALEIRAS-COUTO M M 等[32]从叶片、发酵前的葡萄汁和葡萄酒中提取的 DNA 均扩增出了叶绿体 SSR 位点，为叶绿体 SSR 标记在葡萄酒中葡萄品种的鉴定提供重要参考。

#### 2.2.2 DNA 条形码标记

基于叶绿体基因组的单基因序列、单基因间隔序列或组合序列，具有操作简单、重复性好等优点。如 *matK*[33, 34]、*rbcL*[35, 36]、*trnH-psbA*[37, 38]等已被广泛应用于物种鉴定。范豫杰等[39]利用

*matK+rbcL+trnH-psbA* 的组合 DNA 条形码，可以鉴别虞美人 *Papaver rhoeas* 和罂粟 *P. somniferum*。李冉郡等[40]基于叶绿体基因组比较分析，筛选并验证 *rps16-trnQ*、*psaA-ycf3*、*psbE-petL*、*ndhF-rpl32* 与 *trnT-trnL* 片段均可作为特异 DNA 条形码，用于鉴定唐古特大黄 *Rheum tanguticum*、药用大黄 *R. officinale* 和掌叶大黄 *R. palmatum* 及其混伪品，其中扩增长度最短的 *ndhF-rpl32* 片段还适用于大黄全产业链产品的基原鉴定。DNA 条形码也被用于某些种下水平的鉴别，如王馨等[41]利用筛选出的北柴胡特异 DNA 条形码 *petN-psbM* 和 *ndhF-rpl32* 进行联合分析，结果显示北柴胡的 6 个产地均拥有特异单倍型，使用这对联合标记为鉴别不同产地的北柴胡种质资源奠定基础。

但这些 DNA 条形码序列也存在变异位点少、通用性差等问题。相比于这些序列片段，叶绿体全基因组包含更丰富的变异位点。在植物系统发育研究和物种鉴定中直接利用叶绿体全基因组作为超级条形码，可较大地提高复杂属的物种分辨率[42, 43]，如使用叶绿体全基因组可鉴别贝母属 *Fritillaria* 的多个物种（包括 2 个存疑种）[44, 45]。ZHANG W 等[46]对 4 个通用 DNA 条形码、6 个水稻物种特异 DNA 条形码和叶绿体全基因组的稻属 *Oryza* 物种鉴别能力进行评估，叶绿体基因组表现出最高的分辨力，可以准确鉴别稻属所有物种。张召雷[47]利用叶绿体基因组及线粒体基因组可以鉴别铁笛丸的植物、真菌和动物成分。

#### 2.2.3 DSS 标记

在动物的物种鉴别中，利用 COI 序列即可对大部分物种进行区别。而不同科属植物的特异 DNA 条形码通常不同，无法通过一种通用 DNA 条形码即完成对大多数物种的鉴别[48, 49]。现有的分子标记通常无法兼顾通用性和准确性，利用 CGIR 数据开发的新型分子标记 DSS 很好地弥补了这一局限性。利用 DSS 标记建立的 DNA 分子鉴定方法不依赖 DNA 序列相似度，其特异性可达 100%，且通用性为 79.38%，高于 *matK*、*ycf5*、*ycf1*、*atpB*。

利用 DSS 标记鉴定的准确性已在人参和西洋参、苍术、白术和关苍术等的鉴定中得到充分验证[16, 50]。2022 年版《安徽省中药材标准》也已收载凤丹皮、刺梨、木槿皮、土人参、岗梅根、南瓜蒂、赶黄草、椒目（花椒）8 种药材的 DSS 分子鉴别方法[51]。DSS 标记还可以与 Sanger 测序、特异性 PCR、高通量测序等不同方法结合，适用于混合样品的植物成分鉴定、生物多样性调查、濒危物种贸易监督等应用场景。基于 DSS 标记和等位基因特异性 PCR（Allele specific PCR, AS-PCR）技术，已成功建立九里香[52]、木槿皮[53]、中亚苦蒿[54]、黄芪[55]、地骨皮[56]的位点特异性 PCR 鉴别方法。联合 DSS 标记和限制性片段长度多态性（Restriction fragment length polymorphism, RFLP），已成功建立香加皮与五加皮、地骨皮的 PCR-RFLP 鉴别方法[57]。由于 DSS 标记为 40 bp 的片段，其在存在严重 DNA 降解的植物源性产品的全产业链产品鉴定中具有广泛前景和发展空间。

CGIR 提供基于叶绿体基因组开发的 3 种常用分子标记 SSR、DNA 条形码和 DSS 的序列信息及其电子 PCR 引物，其收录条目数远超过现有的其他叶绿体基因组分子标记数据库。CGIR 同时提供了物种鉴定正反向查询工具，极大地方便了研究人员利用叶绿体基因组进行分子鉴定标记开发。

### 2.3 叶绿体基因工程

叶绿体表达系统具有原核表达特性，拥有丰富的代谢模块[58]，在同质化的转叶绿体细胞中，外源基因可多达 10 000 个拷贝[59]，能够实现外源基因的高效表达。此外，叶绿体基因组的单系遗传能够有效降低外源基因通过花粉逃逸的可能性[60]。这些特性使叶绿体成为基因工程的理想受体[61]，早期的叶绿体基因工程主要集中在增强作物产量、品质和抗逆性相关靶基因的表达水平。KUMAR S 等[62]利用非绿色组织作为外源 DNA 受体，通过体细胞胚发生实现叶绿体转化，将甜菜碱脱氢酶的编码基因 BADH 导入胡萝卜叶绿体基因组中。与未转化细胞相比，转基因胡萝卜细胞的 BADH 酶活性提高了 8 倍，生长速度提高了 7 倍，甜菜碱积累量增加了 50~54 倍，明显提高了转基因胡萝卜植物的耐盐性。近年来，研究人员也开始尝试下调特定靶基因表达水平的新策略，如 JIN S 等[63]通过叶绿体基因组表达鳞翅目几丁质合成酶、细胞色素 p450 单加氧酶和 V-ATP 酶的 dsRNAs，啃食这些叶片则叶绿体源性的 dsRNAs 将抑制昆虫生存所需的这三种必需蛋白质的转录，从而显著降低了幼虫的净重、生长和化蛹率。除此之外，叶绿体基因工程在生产生长激素[64]、蛋白质药物[65]和疫苗[66]上也具有广阔

的应用前景。

叶绿体基因组工程是通过将外源基因整合到基因间间隔区域而不破坏原生叶绿体基因来完成的，同时外源基因的表达受叶绿体基因启动子和非翻译区的调控。因此，叶绿体基因组序列的揭示对叶绿体转化载体的构建是十分重要的，CGIR 通过对叶绿体基因组的基因信息和功能注释进行校正，为叶绿体基因工程提供了叶绿体基因组结构等重要信息。

## 3 展望

### 3.1 叶绿体基因组组装和注释的标准化

在叶绿体基因组研究迅速与蓬勃发展的当今，叶绿体基因组的正确组装和准确注释仍是影响其是否被有效利用的重要因素。大多数陆地植物的叶绿体基因组为环状四分体结构，而少数物种则具有特殊结构，如伞藻 *Acetabularia cliftonii* 的叶绿体基因组为线状结构[67]；部分石松类植物的两个重复区为正向重复[68]；鹰嘴豆 *Cicer arietinum* 等部分豆科植物则丢失了 1 个重复区[69]。在具有经典四分体结构的叶绿体基因组中，含有长反向重复区的叶绿体基因组普遍同时存在单拷贝区域方向不同的 2 种结构单倍型[70]。叶绿体基因组的这种结构异质性可能会使单拷贝区域的组装方向存在差异，从而影响叶绿体基因组比对的结果。如早先的研究中认为海岛棉 *Gossypium barbadense* 与陆地棉 *G. hirsutum* 的叶绿体基因组小单拷贝区方向不同，存在结构差异[71]。因此，CGIR 在科级分类单元，选取模式属模式种的叶绿体基因组作为参考序列，校正其他叶绿体基因组组装方向，以减少叶绿体基因组结构异质性对后续分析的影响。其收录的数据不仅可为其他叶绿体基因组的分析利用提供可靠的参考序列，相应的审编和注释流程也为建立叶绿体基因组的组装、注释和分析标准化流程提供重要参考[72]。

### 3.2 助力中药资源、中药鉴定、中药活性成分生物智造研究

叶绿体基因组是进行中药分子鉴定、中药新资源发现与评价、药用资源系统发育学、药用植物叶绿体基因组工程等研究的重要数据基础。虽然随着测序技术的发展，叶绿体基因组测序数量已有大幅增长。但相对于其他农作物，药用植物的叶绿体基因组在其中的占比仍较低。已发表数据大部分来源于常用大宗药材基原植物，这限制了叶绿体基因组在中药领域的深入研究。CGIR 收录了目前全球物种数量最多的叶绿体基因组及其 DSS、SSR 分子标记、DNA 条形码。并依托第四次全国中药资源普查建立的国内最大中药资源标本库，发布了团队自测的 721 种未发表的药用植物叶绿体基因组。同时，CGIR 还收录了 2020 年版《中国药典》和《全国中草药汇编》记载的相关药用信息，为提升药用植物叶绿体基因组利用能力提供有效支撑。

### 3.3 建立持续更新的叶绿体基因组数据库

随着新技术、新仪器、新算法的出现，新认知和新数据将不断涌现，这将给各个数据库的维护、管理和使用带来更大的挑战。CGIR 作为当前全球收录物种数量最多叶绿体基因组数据库，为用户提供了全面且可靠的数据信息，也在很大程度上为用户提供了一个叶绿体基因组数据共享平台。自 2022 年发布至今，CGIR 已进行了 2 次数据更新。为了反映叶绿体基因组的最新研究成果并进一步满足用户需求，未来 CGIR 还将保持数据库的不断维护与更新、升级与改造，以方便全世界科学工作者的使用。

## [参考文献]

[1] REYES-PRIETO A, WEBER A P, BHATTACHARYA D. The origin and establishment of the plastid in algae and plants[J]. Annu Rev Genet, 2007, 41:147.
[2] DOBROGOJSKI J, ADAMIEC M, LUCIŃSKI R. The chloroplast genome: a review[J]. Acta Physiol Plant, 2020, 42:98.
[3] BOBIK K, BURCH-SMITH T M. Chloroplast signaling within, between and beyond cells[J]. Front Plant Sci, 2015, 6:781.
[4] SHINOZAKI K, OHME M, TANAKA M, et al. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression[J]. EMBO J, 1986, 5(9):2043.
[5] SINGH B P, KUMAR A, KAUR H, et al. CpGDB: a comprehensive database of chloroplast genomes[J]. Bioinformation, 2020, 16(02):171.
[6] LIU T, CUI Y, JI X, et al. OGDA: a comprehensive organelle genome database for algae[J]. Database, 2020, 2020:baaa097.
[7] RUI ZHANG, FANGFANG GE, HUAYANG LI, et al. PCIR: a database of plant chloroplast inverted repeats[J]. Database, 2019, 2019:baz127.
[8] SABLOK G, PADMA RAJU G V, MUDUNURI S B, et al. ChloroMitoSSRDB 2.00: more genomes, more repeats, unifying SSRs search patterns and on-the-fly repeat detection[J]. Database, 2015, 2015:bav084.

[9] LOCATELLI N S, MCINTYRE P B, THERKILDSEN N O, et al. GenBank's reliability is uncertain for biodiversity researchers seeking species-level assignment for eDNA[J]. Proc Natl Acad Sci USA, 2020, 117(51):32211.
[10] ABEYSOORIYA M, SORIA M, KASU M S, et al. Gene name errors: lessons not learned[J]. PLoS Comput. Biol, 2021, 17(7):e1008984.
[11] CHEN M, MA Y, WU S, et al. Genome Warehouse: a public repository housing genome-scale data[J]. Genomics Proteomics Bioinformatics, 2021, 19(4):584.
[12] CNCB-NGDC MEMBERS AND PARTNERS. Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2024[J]. Nucleic Acids Res, 2024, 52(D1):D18.
[13] HEBERT P D, CYWINSKA A, BALL S L, et al. Biological identifications through DNA barcodes[J]. Proc Biol Sci, 2003, 270(1512):313.
[14] HEBERT P D, GREGORY T R. The promise of DNA barcoding for taxonomy[J]. Syst Biol, 2005, 54(5):852.
[15] YU J, WU X, LIU C, et al. Progress in the use of DNA barcodes in the identification and classification of medicinal plants[J]. Ecotoxicol Environ Saf, 2021, 208:111691.
[16] HUA Z, JIANG C, SONG S, et al. Accurate identification of taxon-specific molecular markers in plants based on DNA signature sequence[J]. Mol Ecol Resour, 2023, 23(1):106.
[17] ZHANG Q, LIU Y, SODMERGEN. Examination of the cytoplasmic DNA in male reproductive cells to determine the potential for cytoplasmic inheritance in 295 angiosperm species[J]. Plant Cell Physiol,2003, 44(9):941.
[18] 陈超南， 陆嘉惠， 李学禹， 等． 甘草属种间杂交种叶绿体 DNA 父系遗传的发现及分析[J]． 广西植物， 2017， 37（2）：162．
[19] WOLFE K H, LI W H, SHARP P M. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs[J]. Proc Natl Acad Sci USA, 1987, 84(24):9054.
[20] 樊守金， 郭秀秀． 植物叶绿体基因组研究及应用进展[J]． 山东师范大学学报（自然科学版）， 2022， 37（1）：22．
[21] LI H T, YI T S, GAO L M, et al. Origin of angiosperms and the puzzle of the Jurassic gap[J]. Nat Plants, 2019, 5(5):461.
[22] MENG W Q, LIU K, HAN X N. Insights into phylogenetic analyses of Amaryllidaceae based on complete chloroplast genome of the endemic Chinese genus *Shoubiaonia* sp.[J]. Sci Hortic, 2024, 337:113515.
[23] NIU T, TIAN C, YANG Y, et al. Complete chloroplast genome of *Corethrodendron fruticosum* (Papilionoideae: Fabaceae): comparative and phylogenetic analysis[J]. Genes, 2023, 14(6):1289.
[24] LI T, ZHANG S, DENG Y, et al. Comparative analysis of chloroplast genomes for the genus *Manglietia* Blume (Magnoliaceae): molecular structure and phylogenetic evolution[J]. Genes, 2024, 15(4):406.
[25] 罗晗睿， 王罗云， 张建国， 等． 胡颓子科叶绿体基因组系统发育分析与演化趋势推断[J]． 西北植物学报， 2024， 44（6）：891．
[26] QI K, CHEN Z, LI X, et al. Complete genome and comprehensive analysis of *Knorringia sibirica* chloroplast[J]. Horticulturae, 2024, 10(3):268.
[27] SHI W B, SONG W C, LIU J, et al. Comparative chloroplast genome analysis of *Citrus* (Rutaceae) species: insights into genomic characterization, phylogenetic relationships, and discrimination of subgenera[J]. Sci Hortic, 2023, 313:111909.
[28] ZENG Q, CHEN M, WANG S, et al. Comparative and phylogenetic analyses of the chloroplast genome reveal the taxonomy of the *Morus* genus[J]. Front. Plant Sci, 2022, 13:1047592.
[29] 王化坤， 娄晓鸣， 章镇． 叶绿体微卫星在植物种质资源研究中的应用[J]． 分子植物育种， 2006，4（S1）：92．
[30] 李思巧， 韦伊， 刘洪妤， 等． 花椒 cpSSR 标记开发及在种间、种内的通用性分析[J]． 浙江农林大学学报， 2019， 36（6）：1241．
[31] 严卓彦， 夏瑛瑛， 崔洁， 等． 基于 DNA 条形码及 SSR 技术鉴定风藤和山荪[J]． 现代中药研究与实践， 2023， 37（2）：23．
[32] BALEIRAS-COUTO M M, EIRAS-DIAS J E. Detection and identification of grape varieties in must and wine using nuclear and chloroplast microsatellite markers[J]. Anal Chim Acta, 2006, 563(1/2):283.
[33] 赵凯辉， 袁芳， 赵芳玉， 等． 基于 *ITS2* 和 *matK* 序列的红景天属植物条形码鉴定研究[J]． 高原农业， 2022， 6（5）：462．
[34] 徐林， 赵芳， 杨贵清， 等． *matK* 序列在绞股蓝及其混伪品中的应用[J]． 安徽农学通报， 2023， 29（7）：22．
[35] 夏若成， 张晓春， 王笑笑， 等． 基于 *rbcL* 序列的大麻鉴定[J]． 法医学杂志， 2021， 37（2）：187．
[36] 孙思雅， 陈云， 王琪瑞， 等． 基于 *ITS2* 和 *rbcL* 条形码的金钱草及其地方习用品与混伪品分子鉴定技术研究[J]． 中药材，2020， 43（6）：1336．
[37] 郭慧． 酸模属 DNA 条形码的筛选与鉴定研究[D]． 哈尔滨：黑龙江中医药大学， 2017．
[38] 侯静， 黄勇， 黎理， 等． 广西斑地锦及其混淆品 DNA 条形码分子鉴定[J]． 湖北农业科学， 2022， 61（10）：156．
[39] 范豫杰， 黄磊， 苏世达， 等． 基于 *matK+rbcL+trnH-psbA* 联合分析法对罂粟属的植物鉴定[J]． 昆明医科大学学报， 2023，44（1）：12．
[40] 李冉郡． 基于 DNA 条形码的大黄全产业链基原物种鉴定研究[D]． 重庆：西南交通大学， 2022．
[41] 王馨， 尹光耀， 张志飞， 等． 北柴胡特异 DNA 条形码筛选及种质资源鉴定[J]． 中草药， 2023， 54（17）：5703．
[42] KANE N C, CRONK Q. Botany without borders: barcoding in focus[J]. Mol Ecol, 2008, 17(24):5175.
[43] LI X, YANG Y, HENRY R J, et al. Plant DNA barcoding: from gene to genome[J]. Biol Rev Camb Philos Soc, 2015, 90(1):157.
[44] CHEN Q, HU H, ZHANG D. DNA barcoding and phylogenomic analysis of the genus *Fritillaria* in China based on complete chloroplast genomes[J]. Front Plant Sci, 2022, 13:764255.
[45] LIANG Y Z, ZHANG J, WANG X, et al. Applying DNA barcoding to identify the cultivated provenance of *Fritillaria taipaiensis* P. Y. Li and its related species[J]. J Appl Res Med Aromat Plants, 2024, 39:100530.
[46] ZHANG W, SUN Y, LIU J, et al. DNA barcoding of *Oryza*: conventional, specific, and super barcodes[J]. Plant Mol Biol, 2021, 105(3):215.
[47] 张召雷． 基于叶绿体及线粒体基因组全长的中成药铁笛丸生物成分鉴定方法构建[D]． 承德：承德医学院， 2023．
[48] 王丽， 赵锐， 张秋芳． DNA 条形码分子鉴定技术在中药材鉴定中的应用[J]．食品与药品， 2022， 24（4）：388．
[49] 黄璐琦， 袁媛， 袁庆军， 等． 中药分子鉴定发展中的若干问题探讨[J]． 中国中药杂志， 2014， 39（19）：3663．
[50] 王多梅， 蒋超， 蒲婧哲， 等． 《中国药典》植物药基原鉴定 DSS 标准序列库的建立[J/OL]． 中国中药杂志， 2024．
[51] 安徽省药品监督管理局． 安徽省中药材标准[M]． 北京：中国医药科技出版社， 2022．
[52] 陈梓媛， 赵玉洋， 谢旭桃， 等． 多基原九里香药材的多重位点特异性 PCR 鉴别[J]． 中国实验方剂学杂志， 2022， 28（17）：106．

[53] 刘亚男， 华中一， 赵玉洋， 等. 基于 DSS 标记特异性 PCR 鉴别冷背药材木槿皮基原植物及其混伪品[J]. 中国实验方剂学杂志， 2022， 28（17）：28.
[54] 刘志浩， 陈梓媛， 李晓琳， 等. 蒿属外来药用资源中亚苦蒿的特异性 PCR 鉴别[J]. 中国实验方剂学杂志， 2022， 28（17）：127.
[55] 罗丽， 胡力， 蒋超， 等. 蒙古黄芪、膜荚黄芪及混伪品种子的位点特异性 PCR 鉴别[J]. 中国实验方剂学杂志， 2024， 30（4）：21.
[56] 石旸， 胡力， 赵玉洋， 等. 地骨皮的多重位点特异性 PCR 鉴别[J]. 中国实验方剂学杂志， 2024， 30（4）：35.
[57] 刁晓微， 刘亚男， 金艳， 等. 香加皮与五加皮、地骨皮的 PCR-RFLP 鉴别[J]. 中国实验方剂学杂志， 2024， 30（4）：42.
[58] BOEHM C R, BOCK R. Recent advances and current challenges in synthetic biology of the plastid genetic system and metabolism[J]. Plant Physiol, 2019, 179(3):794.
[59] JIN S, DANIELL H. The engineered chloroplast genome just got smarter[J]. Trends Plant Sci, 2015, 20(10):622.
[60] DANIELL H. Transgene containment by maternal inheritance: effective or elusive?[J]. Proc Natl Acad Sci U S A, 2007, 104(17):6879.
[61] DANIELL H, LIN C S, YU M, et al. Chloroplast genomes: diversity, evolution, and applications in genetic engineering[J]. Genome Biol, 2016, 17(1):134.
[62] KUMAR S, DHINGRA A, DANIELL H. Plastid-expressed betaine aldehyde dehydrogenase gene in carrot cultured cells, roots, and leaves confers enhanced salt tolerance[J]. Plant Physiol, 2004, 136(1):2843.
[63] JIN S, SINGH N D, LI L, et al. Engineered chloroplast dsRNA silences cytochrome p450 monooxygenase, V-ATPase and chitin synthase genes in the insect gut and disrupts *Helicoverpa zea* larval development and pupation[J].Plant Biotechnol J, 2015, 13(3):435.
[64] JIN S, KANAGARAJ A, VERMA D, et al. Release of hormones from conjugates: chloroplast expression of beta-glucosidase results in elevated phytohormone levels associated with significant increase in biomass and protection from aphids or whiteflies conferred by sucrose esters[J]. Plant Physiol, 2011, 155(1):222.
[65] SU J, ZHU L, SHERMAN A, et al. Low cost industrial production of coagulation factor Ⅸ bioencapsulated in lettuce cells for oral tolerance induction in hemophilia B[J]. Biomaterials, 2015, 70:84.
[66] DAVOODI-SEMIROMI A, SCHREIBER M, NALAPALLI S, et al. Chloroplast-derived vaccine antigens confer dual immunity against cholera and malaria by oral or injectable delivery[J]. Plant Biotechnol J, 2010, 8(2):223.
[67] GREEN B R. Covalently closed minicircular DNA associated with *Acetabularia* chloroplasts[J]. Biochim Biophys Acta, 1976, 447(2):156.
[68] MOWER J P, MA P F, GREWE F, et al. Lycophyte plastid genomics: extreme variation in GC, gene and intron content and multiple inversions between a direct and inverted orientation of the rRNA repeat[J]. New Phytol, 2019, 222(2):1061.
[69] JANSEN R K, WOJCIECHOWSKI M F, SANNIYASI E, et al. Complete plastid genome sequence of the chickpea (*Cicer arietinum*) and the phylogenetic distribution of *rps12* and *clpP* intron losses among legumes (Leguminosae)[J]. Mol Phylogenet Evol, 2008, 48(3):1204.
[70] WANG W W, LANFEAR R. Long-reads reveal that the chloroplast genome exists in two distinct versions in most plants[J]. Genome Biol Evol, 2019, 11(12):3372.
[71] IBRAHIM R I, AZUMA J, SAKAMOTO M. Complete nucleotide sequence of the cotton (*Gossypium barbadense* L.) chloroplast genome with a comparative analysis of sequences among 9 dicot plants[J]. Genes Genet Syst, 2006, 81(5):311.
[72] QU X J, ZOU D, ZHANG R Y, et al. Progress, challenge and prospect of plant plastome annotation[J]. Front Plant Sci, 2023, 14:1166140.