时珍法物种鉴定原理与研究策略

甘雨桐[1], 辛天怡[1,2], 许文杰[1,2], 郝利军[1], 齐桂红[1], 娄　千[1], 宋经元[1,2,3*]

(1. 中国医学科学院、北京协和医学院药用植物研究所, 国家中医药管理局中药资源保护重点研究室, 北京 100193; 2. 中药资源教育部工程研究中心, 北京 100193; 3. 中国医学科学院、北京协和医学院药用植物研究所云南分所, 云南 景洪 666100)

摘要: 天然药物的主要来源包括植物、动物和微生物等生物物种。这些物种的准确鉴定是天然药物研发的基础。本文提出一种全新的物种鉴定方法: 时珍法 (analysis of whole-genome, AGE), 即从物种全基因组寻找特异靶标序列, 精准识别特异靶标序列, 实现物种准确鉴定的分子诊断方法。笔者详细阐述了时珍法利用不同物种基因组序列必然存在差异的原理实施物种鉴定, 提出该方法的实施策略可分为研究和应用两个层面, 分析其特点评判该方法具有原理可靠、特异性强、适用性广等优势, 进一步论述了在体系构建过程涉及的基因组获取、生物信息分析、数据库构建等三个关键问题。综上所述, 本文为生物信息软件和商品试剂盒的后续开发提供理论基础和方法指导, 表明时珍法在不同对象、不同学科、不同行业等方面具有巨大应用潜力。

关键词: 时珍法; 物种鉴定; 全基因组; 生物信息分析; 数据库构建

中图分类号: R932　　**文献标识码**: A　　**文章编号**: 0513-4870(2023)08-2364-11

Principles and strategies for species identification based on analysis of whole-genome

GAN Yu-tong[1], XIN Tian-yi[1,2], XU Wen-jie[1,2], HAO Li-jun[1], QI Gui-hong[1], LOU Qian[1], SONG Jing-yuan[1,2,3*]

(*1. Key Lab of Chinese Medicine Resources Conservation, State Administration of Traditional Chinese Medicine of the People's Republic of China, Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100193, China; 2. Engineering Research Center of Chinese Medicine Resource, Ministry of Education, Beijing 100193, China; 3. Yunnan Branch, Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences and Peking Union Medical College, Jinghong 666100, China*)

Abstract: The main sources of natural drugs include various biological species such as plants, animals, and microorganisms. The accurate identification of these species is the bedrock of natural drug development. We propose a novel method of species identification in this paper: analysis of whole-genome (AGE), a molecular diagnostic method used to identify species by finding species-specific sequences from the whole genome and precisely recognizing the specific target sequences. We elaborate that the principle for species identification based on AGE is that the genome sequences of diverse species must differ and divide the implementation strategy of the method into two levels of research and application. Based on our analysis of its characteristics, the method would have the potential advantages of reliable principle, high specificity, and wide applicability. Moreover, three crucial concerns related to building method systems including genome acquisition, bioinformatics analysis, and database construction, are further discussed. In summary, we offer theoretical underpinnings and methodological guidance

收稿日期: 2023-04-26;　修回日期: 2023-06-27.
基金项目: 国家自然科学基金资助项目 (81874339); 中国医学科学院医学与健康科技创新工程 (2022-I2M-2-001).
*通讯作者 Tel: 86-10-57833199, E-mail: jysong@implad.ac.cn
DOI: 10.16438/j.0513-4870.2023-0513

for the development of bioinformatics software and commercial kits, indicating AGE has great application potential in objects, subjects, and industries.

Key words: analysis of whole-genome; species identification; whole genome; bioinformatics analysis; database construction

物种鉴定是药学、中药学、生物科学、环境科学、食品科学等学科的基础, 在自然科学等领域具有重要意义。在药学研究领域中, 物种的准确鉴定与人类健康密切相关, 应用广泛, 为药品研究、生产、销售和管理等提供基础技术支持和保障。植物、动物和微生物是天然药物的主要来源, 但过度采集和利用会对物种野生资源造成威胁, 通过物种鉴定, 有助于保护物种资源, 维护生态平衡。在药品生产销售过程中, 物种鉴定可以帮助实现溯源管理, 提高生产销售透明度和可信度。中国药典是确保药品质量和安全性的重要标准, 物种鉴定作为一种基础技术方法, 可为中国药典标准修订提供依据。

物种的准确鉴定和正确分类也是保护生物多样性的重要手段[1], 生物多样性是保持生态系统稳定的关键[2,3]。为防止全球生物多样性的持续下降和保护生态环境稳定, 对所有物种基因组进行测序是突破口。2018年4月, 多国科学家共同提出“地球生物基因组计划”——在十年内成功组装并注释地球上近150万种已知真核生物物种的基因组[4], 其规模比人类基因组计划更宏大, 被誉为“生物登月计划”。2021年Science公布的125个全世界最前沿问题中, 多数生命科学领域问题都可以在物种基因组中找到答案。2022年诺贝尔生理学或医学奖颁给瑞典科学家Svante Pääbo, 以表彰他在古基因组学和揭示人类起源方面所做出的卓越贡献。除上述研究外, 基因组学也赋予中药研究新的活力, 衍生的本草基因组学在基因组水平上阐明中药作用的分子机制, 为中药创新研发提供动力[5]。基因组学是生命科学研究皇冠上一颗闪亮的明珠。据此, 本文提出一种全新的物种鉴定方法: 时珍法, 又称全基因组分析法或阿哥法。与DNA条形码以单一或多条小片段基因序列给全球的物种编码不同, 本文介绍的时珍法利用物种基因组全部遗传信息对生物进行鉴定。以基因组作为支撑, 时珍法具有鉴定任意物种的潜力, 包括一直被认为是鉴定难点的近缘物种。1977年, Sanger等[6]完成首个全基因组测序, 噬菌体φX174基因组, 大小为5.836 kb。“地球生物基因组计划”预计到2022年底完成测序、组装和注释3 000多个物种基因组[7]。根据美国国家生物技术信息中心(National Center for Biotechnology Information, NCBI)网站统计, 截至2022年8月, 共公开24 955个真核生物基因组。同时, 测序技术经过不断变革, 其费用也呈断崖式下降。根据美国人类基因组研究所(National Human Genome Research Institute Home, NHGRI)追踪的数据(图1), 2021年完成单人基因组测序仅需562美元, 约为2001年花费的近十七万分之一! 如果把接连不断的生物基因组揭秘和蓬勃发展的生物信息学视作时珍法大步前进的左右腿, 那么可通过多种实验技术进行鉴定这一特点就如同一双舒适的鞋子, 共同支撑时珍法提出物种鉴定领域新视角。

本文详细阐述时珍法的定义和原理, 从方法的实施策略、技术路线、特点以及实施过程的关键问题等方面进行论述, 提出建立完整时珍法体系的初步方案, 以供讨论和商榷。

1 时珍法简介

1.1 时珍法的定义

时珍法是从物种全基因组寻找特异靶标序列, 精准识别特异靶标序列, 实现物种准确鉴定的分子诊断方法。

1.2 时珍法的原理

时珍法的基本原理是不同物种的全基因组序列必然存在差异。全基因组涵盖物种所有遗传信息, 也就是全部核酸序列。正如世界上没有完全相同的两片树叶, 不同物种在基因组层面存在差异。将一个物种基因组同其他物种基因组进行分析比对, 能够找到识别此物种的特异靶标序列。时珍法基于生物信息学分析和实验操作两个步骤找到目标物种的特异靶标序列, 并通过识别特异靶标序列进行物种鉴定。

2 时珍法物种鉴定的实施策略

时珍法物种鉴定的实施策略分为研究和应用两个层面。研究层面分为两步: 首先, 利用生物信息技术从数据层面找到物种特异靶标序列; 其次, 利用不同实验技术从实验层面识别样品中目标物种特异靶标序列, 从而实现物种鉴定。而在方法应用层面, 则无需进行生物信息分析, 直接从研究层面建成的特异靶标序列库中选取合适的特异靶标进行实验即可。下文分别对生物信息分析和实验技术进行具体阐述。

2.1 生物信息分析

生物信息分析主要分为三部分(图2[8])。第一步工作即收集目标物种与其他物种的基因组, 再基于测序、

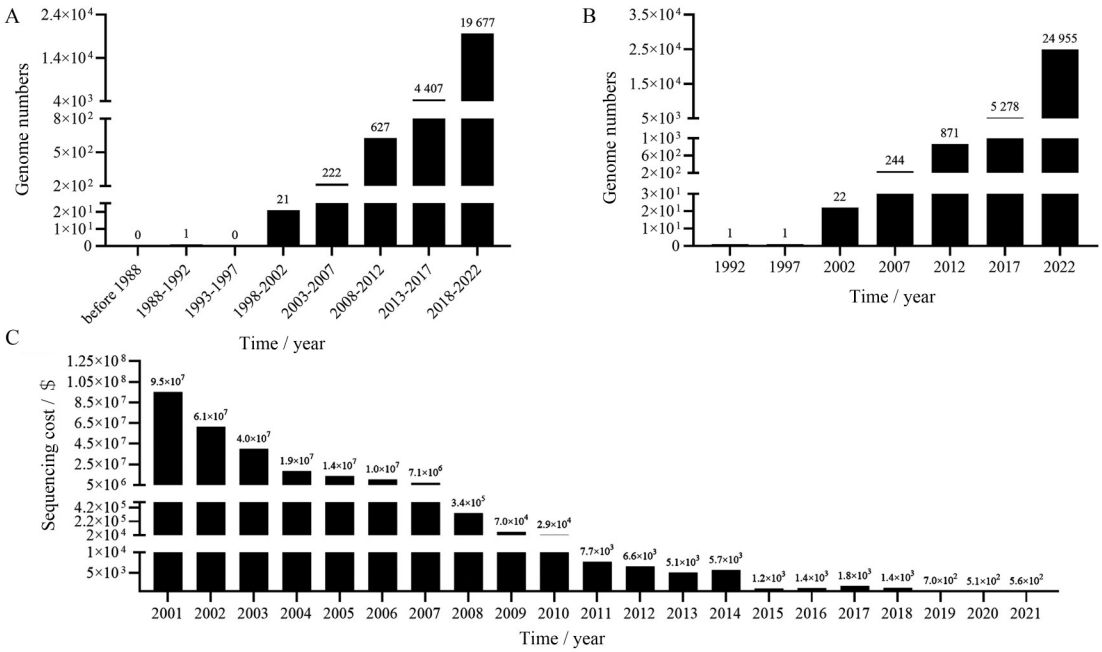

Figure 1 Genome disclosure and genome sequencing cost survey statistics. A: According to NCBI statistics, changes in the number of publicly available eukaryotic genomes every 5 years; B: According to NCBI statistics, changes in the total number of publicly available eukaryotic genomes in key years; C: According to NCBI statistics, changes in human genome sequencing costs between 2001 and 2021

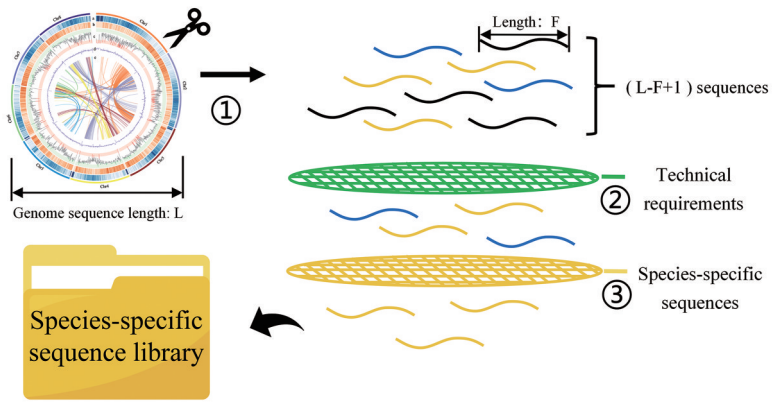

Figure 2 Pattern diagram for bioinformatics analysis of the whole genome of species. Based on the subsequent assay purpose, requirements and genome annotation as well as sequence information of the target species and other species, analysis of whole-genome (AGE) constructs a fragment sequence library of the species by screening. The sequences in the target species fragment library are compared with the fragment sequence libraries of other species, and the specific fragment sequences that exist only in the target species are retained to construct the specific target sequence library of the target species. The species identification is based on this specific target sequence library. Genome schematic diagram was reprinted with permission from reference[8], copyright (2022) John Wiley and Sons. Folder image: Freepik.com

基因组编辑[9,10]、微滴式数字 PCR (droplet digital PCR, ddPCR)[11]、实时荧光定量 PCR (quantitative real-time PCR, qPCR)[12,13]、DNA 芯片[14]等不同实验技术对于序列长度要求, 将所有物种全基因组序列分成 (L-F+1) 条长度为 F 的片段序列 (L 表示全基因组序列长度, F 表示文库中片段序列长度), 构建供不同实验技术使用的物种片段序列库。某些实验技术除序列长度外, 还对片段序列有特殊要求, 如: 应用 CRISPR-Cas12a 技术时, 序列需包含 PAM (TTTV) 序列。第二步工作则从物种片段序列库中筛选符合实验技术要求的序列, 组成物种候选片段序列库。第三步将目标物种候选片段序列库中每一条序列同其他物种候选片段序列库中所有序列进行比对, 剔除一致或同源性较高的片段序列, 保留仅存在于目标物种全基因组的特异片段序列, 构建目标物种特异靶标序列库。其中第二步不是必需步骤。重复以上三步操作, 即可得到以实验技术为分类

标准的目标物种集成特异靶标序列库。

以基因组编辑CRISPR-Cas12a体系作为实施案例, 进行具体阐述。第一步从所有物种的全基因组中随机提取长度为25 bp的片段序列组成片段序列库; 第二步提取包含PAM的序列构建候选片段序列库; 第三步将目标物种的候选靶标序列与其他物种片段序列库进行比对, 挑选除PAM序列外21 bp中存在3个及以上碱基错配或插入/缺失的序列, 构建目标物种的特异靶标序列库。构建测序、qPCR、ddPCR、DNA芯片、琼脂糖凝胶等其他实验技术的物种特异靶标序列库时, 根据方法对序列的要求, 按照上述生物信息分析三步工作流程进行即可。

2.2 实验技术操作策略

不同实验技术具有不同技术特点, 因此可根据不同的鉴定需求和应用场景选择不同实验技术 (表1)。例如, Sanger测序技术可直接获得序列结果, 基因组编辑技术则可以避免PCR扩增, 且鉴定结果可通过可视荧光或试纸条呈现, 具有更强的便携性, 满足现场检测的需求。随着后疫情时代的到来, qPCR技术已经得到广泛应用, 具有仪器普及、操作熟练等优势。高通量测序技术能够一次性获取大量序列信息, 完成更全面完整的分析。就像新冠病毒检测分为qPCR和高通量测序, 分别应用于不同检测目的。在这些技术中, 琼脂糖凝胶电泳技术最易操作且最为经济实惠, 但其灵敏度相对较低。高通量测序技术的造价最高, 但灵敏度也最高。因此, 用户可以综合考虑检测的灵敏度需求和经济成本, 选择最适合的实验技术。

由于不同实验技术的操作存在较大差异, 故下文主要根据操作是否需要测序分类阐述7种实验技术的操作策略。每个方法的基本实验操作过程都包括DNA提取。鉴于目前真菌、动物、植物的DNA提取技术皆已发展成熟, 不再加以赘述。

2.2.1 需测序的实验技术操作策略

2.2.1.1 Sanger测序 从目标物种特异靶标序列库挑选特异靶标序列, 并根据特异靶标序列上下游序列设计引物。根据目标产物长度和引物Tm值确定PCR反应条件, 以样品基因组DNA为模板进行PCR扩增, 并对PCR产物进行双向测序。将测序序列去除引物区和低质量区后拼接, 获得待检样品特异靶标序列。如果待检样品特异靶标序列与库中序列片段完全相同, 则判定待检样品与目标物种具有同一性。反之, 则不具有同一性。

2.2.1.2 高通量测序 对待检样品基因组DNA使用Illumina、Nanopore和PacBio等高通量测序技术进行测序, 获得其大量片段序列, 将其与目标物种特异靶标序列进行比对。如果样品的片段序列中包含目标物种特异靶标序列, 则判定待检样品与目标物种具有同一性。反之, 则不具有同一性。

2.2.2 无需测序的实验技术操作策略

2.2.2.1 基因组编辑 (以CRISPR-Cas12a体系为例) 从目标物种特异靶标序列库中随机选择一条靶标序列作为模板合成crRNA, 并根据靶标序列上下游序列设计合成引物。通过常规PCR或者室温扩增对样品DNA进行扩增。Cas12a酶和crRNA结合形成复合物, 与DNA扩增产物结合后, 激活Cas12a酶反式切割活性, 切割带荧光探针的单链DNA (single-stranded DNA, ssDNA)。反应体系有荧光产生, 则判定待检样品与目标物种具有同一性; 反之, 则不具有同一性。当利用试纸条检测时, 若待检样品出现测试条带, 则判定待检样品与目标物种具有同一性; 反之, 则不具有同一性。

2.2.2.2 ddPCR 从目标物种特异靶标序列库中随机选择一条靶标序列合成Taqman荧光探针, 并根据靶标序列上下游序列设计合成特异引物。按比例配制PCR反应体系, 对其进行微滴化处理再放入PCR仪进行反应。结束后, 使用微滴分析仪检测微滴荧光信号。若出现荧光信号, 则判定待检测样品与目标物种具有同一性。反之, 则不具有同一性。

2.2.2.3 qPCR qPCR目前有两种常见的检测方法, 分别是荧光染料嵌合法和荧光探针法。SYBR荧光染料是荧光染料嵌合法最常用的染料。Taqman探针是荧光探针法中的代表性方法。

Table 1 Analysis and comparison of different experimental technologies. * represents non-required; + represents affirmation; - represents negation. The number of + is proportional to the sensitivity

Technology	Sequencing	PCR	Sensitivity	Result	On-site testing
Sanger sequencing	+	+	+	Sequence	-
High-throughput sequencing	+	*	++++	Sequence	-
Genome editing	-	*	++	Fluorescence/test strip	+
ddPCR	-	+	+++	Fluorescence	-
qPCR	-	+	+++	Fluorescence	-
DNA chips	-	+	++	Figure	-
Agarose gel electrophoresis	-	+	+	Figure	-

SYBR荧光染料法主要步骤包括从目标物种特异靶标序列库中随机选择一条靶标序列, 并根据靶标序列上下游序列设计合成引物。将包含待检样品DNA、SYBR染料、引物和ddH_2O等反应体系以及已鉴定为目标物种的标准样品对照反应体系, 放入荧光定量PCR仪中。待反应结束后, 比较待检样品与标准样品的曲线, 如果曲线重合, 则判定待检样品与目标物种具有同一性。反之, 则不具有同一性。

Taqman探针法主要步骤包括从目标物种特异靶标序列库中随机选择一条靶标序列合成Taqman荧光探针, 并根据靶标序列上下游序列设计合成引物。在待检样品DNA扩增体系中加入特异Taqman荧光探针进行实时荧光定量PCR实验。经检测, 体系中有荧光产生, 则判定待检测样品与目标物种具有同一性。反之, 则不具有同一性。由于Taqman探针实验具有兼容多重反应的特性, 因此可以针对同一段扩增序列设计多种探针, 利用不同荧光标记的探针同时检测多个靶标。

2.2.2.4 DNA芯片 从目标物种特异靶标序列库中随机选择靶标序列合成探针固定在DNA芯片上, 并根据靶标序列上下游序列设计合成引物。通过PCR实验, 扩增待检样品并用生物素标记PCR产物。将标记后的PCR产物与含靶标序列探针的DNA芯片在缓冲液中孵育杂交。反应结束后, 使用读取器对结果进行阅读。若在芯片相应位置出现印记, 则判定待检样品与目标物种具有同一性。反之, 则不具有同一性。

2.2.2.5 琼脂糖凝胶电泳 从目标物种特异靶标序列库中随机挑选特异靶标序列, 并根据特异靶标序列上下游序列设计特异性扩增引物。再根据目标产物长度和引物Tm值确定PCR反应条件, 以待检样品和已鉴定为目标物种的标准样品基因组DNA为模板分别进行PCR扩增。对PCR产物进行琼脂糖凝胶电泳检测, 若检测样品出现条带且与目标物种条带位置相同, 则判定待检样品与目标物种具有同一性。反之, 则不具有同一性。

2.3 技术路线

时珍法的完整技术路线见图3, 主要分为生物信息分析和实验操作两部分。建立方法研究体系时, 需按完整的技术路线进行。研究体系成功构建后, 用户使用时珍法鉴定样品时, 只需完成技术路线中的实验操作即可。

3 时珍法与现有分子鉴定方法相比较的特点

形态、显微、理化等传统鉴定方法在物种分类、鉴定中做出了卓越贡献。形态鉴定是目前最基础的鉴定方法,《本草纲目》和《物种起源》都是其代表性成果。但由于物种性状受到环境调控和种内变异的影响, 导致鉴定过程难以顺利进行[15], 因此鉴定过程需要具备丰富的专业知识才能有效支撑[16]。随着人们对生物的认知从形态特征深入到其蕴含的各种物质成分, 理化鉴定逐渐登上历史舞台[17-20]。但理化鉴定在没有标准品的情况下较难进行准确识别, 同时其对材料、仪器的要求较高。

随着分子生物学技术的发展和流行, 顺应人们对更精确的鉴定方法的强烈需求, 分子鉴定方法走入大众视野。蛋白质鉴定大幅度提升鉴定方法灵敏度, 但蛋白质本身不易保存、特异抗体稀缺等难题迫使生物学家找寻新的出路[21,22]。因此直接反映物种间本质差异的DNA分子鉴定技术迅猛发展。DNA性质稳定、来源丰富、便于收集等优点极大程度上助力分子鉴定技术发展。限制性内切酶片段长度多态性[23] (restriction fragment length polymorphism, RFLP)、扩增片段长度多态性[24] (amplified fragment length polymorphism, AFLP)、qPCR[12,13,25]以及DNA芯片[14]等技术都是知名DNA分子鉴定方法。2003年DNA条形码技术面世后[15], 以通用性强、重复性好、操作简单等优点迅速在鉴定领域占领主流地位。DNA条形码技术可有效减轻生物分类学家繁琐鉴定工作面临在物种鉴定方面的巨大负担, 使他们能够将更多精力投入到发现新物种等其他工作[26], 此外, 还可助力中药材质量控制和流通监管, 有效保证公众用药安全[27-31]。但此技术包含测序环节, 使用DNA条形码技术鉴定物种, 检测周期较长。

本文提出的时珍法较其他鉴定方法有以下特点: ① 原理独特, 思路新颖: 时珍法最突出的特点来自物种分类鉴定最底层的原理逻辑——不同物种全基因组基因序列必然存在差异。其他分子鉴定技术都是基于物种中已发表的具有物种特异性的某条片段序列展开研究, 方法之间的最大区别在于采用了新开发的分子生物学相关仪器或技术进行替代。而时珍法跳出在仪器或技术上进行创新的固有思维, 从原理上革新。时珍法最重要的思路是从物种基因组层面进行生物信息分析, 除了分子鉴定普遍应用的序列, 如ITS、COI等, 还能得到大量全新未知特异序列。② 系统集成检测方法, 实践灵活运用: 其他鉴定技术都仅涉及到单种或少数几种技术方法, 而时珍法得益于原理强大, 可以运用前述不同分子生物学技术作为实施手段进行物种鉴定。在方法体系建立过程中, 研究人员以分子鉴定技术作为分类标准, 调整序列筛选标准, 建立在理论层面适用于不同分子鉴定技术的物种集成特异靶标序列库。用户使用方法时可以根据自己鉴定目的、需求, 选

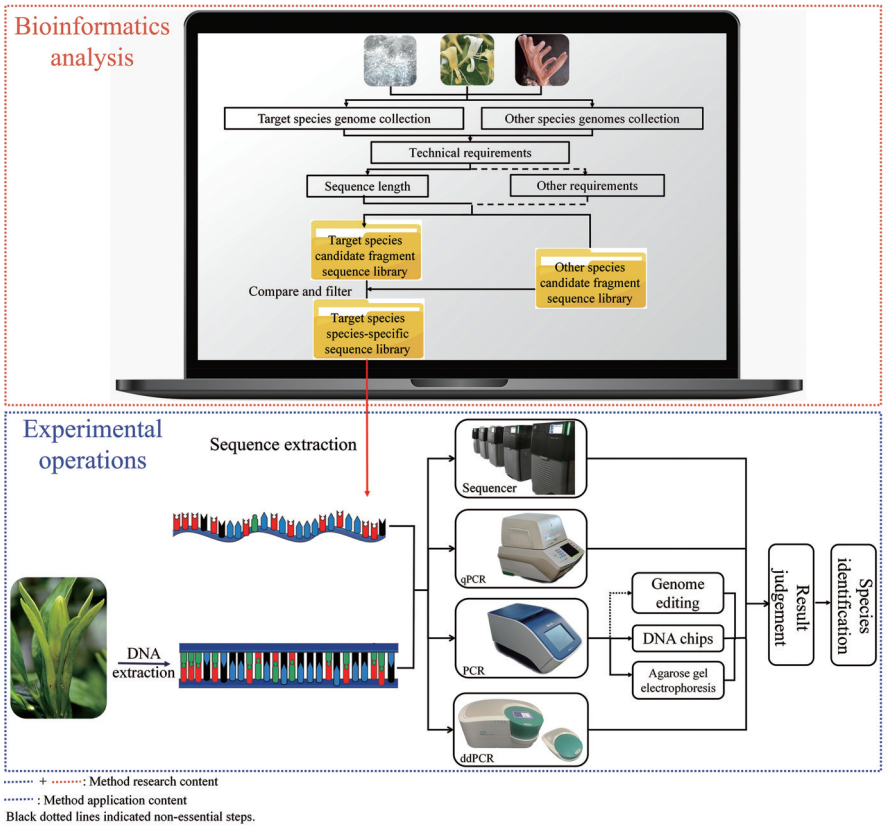

Figure 3 The pipeline of AGE. The technical route of AGE is divided into two parts: bioinformatic analysis and experimental operation. Black dotted lines indicate non-essential steps. The orange dotted box indicates bioinformatics analysis work, and the blue dotted box indicates experimental operations work. Computer image and folder image: Freepik.com. Nucleic acid image: Servier Medical Art Commons Attribution 3.0 Unported License (http://smart.servier.com)

择适宜的分子技术特异靶标序列库中序列, 开展实验。

综上所述, 时珍法兼具以下优点: ① 可靠的底层原理: 在理论上, 通过对目标物种与其相关物种的基因组进行比对分析, 找到目标物种基因组中必然存在的特异序列。② 高特异性: 基于方法的底层逻辑, 对于其他技术难以区分的近缘种、变种和品种甚至有性繁殖个体等对象的鉴定是时珍法的强项。针对上述难以区分的对象, 时珍法直接通过其基因组的序列差异进行鉴定, 所以时珍法也具有高准确性。③ 广适用性: 从本质的序列差别到易于读取结果的试纸条, 该方法的鉴定结果有多种呈现形式。时珍法可以满足使用者在鉴定过程中不同场景的不同需求。当形态特征不明显甚至无形态鉴定特征时, 能使用 AGE 对样品进行鉴定。AGE 也可实现现场检测, 无需在实验室进行复杂操作处理, 显著提升了样品鉴定速度。随着技术发展进步, 时珍法可将新兴科技成果纳为己用, 这赋予时珍法生命和活力, 使其能随着技术革新而不断发展。时珍法被评估为一种能够持续升级更新的成长型鉴定方法。

时珍法可能存在以下缺陷: ① 时珍法旨在通过分析全基因组获得特异性靶标序列实现物种鉴定, 但按照本文阐述的时珍法全基因组分析原理和策略, 无法准确构建生命之树, 也无法深入挖掘物种起源、演化规律和系统分类等信息。通过改变全基因组分析策略, 时珍法实际上也可构建生命之树, 挖掘物种分类演化等信息, 该部分内容不在本文进行具体阐述。② 时珍法前期研究体系建立和数据库构建需要大量资金和人力支撑。但数据库建成后, 时珍法的应用将变得简单便捷。且这一过程能够吸引多学科人才, 为鉴定领域争取更大力度的国家资金支持, 并为中国引领国际科学研究方向提供重要机遇。③ 时珍法无法对不含 DNA 的样品或 DNA 严重降解的样品进行鉴定。

4 构建方法研究体系的关键问题

4.1 基因组的获取

时珍法首要任务是获取目标物种及其相关物种基因组。基因组获取可分为两种情况, 一是物种基因组已公开, 直接联系文章作者或在网站下载即可; 二是物种基因组仍未公开, 可以采用浅层测序技术以获取物种基因组。

公开的基因组数量在近30年呈爆发性增长 (图1)。1992年第一个真核生物基因组公开, 在随后10年, 基因组数量实现从个位数到以十为计量单位的突破。2012年已公开的真核生物基因组数量是2002年的40倍左右。截止到2022年8月, 已公开24 955个真核生物物种基因组, 其中最近5年内公开的真核生物基因组就有近两万个, 占目前已公开基因组数量的80%左右。基因组公开数量正呈井喷式上升。完成地球上所有生物基因组测序应是全球生物学家的共同目标和必要任务。

生物学家预测地球上共有870万左右真核生物物种[32], 目前基因组公开的物种只占极少一部分。如果用户的目标物种和其相关物种目前暂无基因组, 对物种进行浅层测序也能有效获取基因组信息[33]。随着测序技术发展, 基因组测序费用逐年下降。在2021年1 Mb基因组测序费用已经从二十年前5 292.39美元降至0.006美元, 实现88万倍暴跌 (图4)。根据发表的真核生物基因组情况统计, 95%以上核基因组大小集中在1 Mb到3 Gb之间, 95%以上叶绿体基因组大小集中在0.1 Mb到0.17 Mb之间, 95%线粒体基因组集中在0.01 Mb到1 Mb之间。因此现在获取一个物种全基因组30×浅层测序数据一般不会超过540美元, 最低甚至仅需0.18美元。针对大型和超大型基因组, 对其进行浅层测序的费用也在科研经费承受范围之内。例如, 获取9.87 Gb的银杏基因组[34]和146 Gb的重楼百合基因组[35]的30×浅层测序数据的费用是1 776.6美元和26 280美元。基因组的公开数量以指数级速度增长, 同时攻克一个物种基因组的成本越来越低。随着基因组研究的推进, 将来会有更多的物种基因组被公开。由此可见, 在时珍法研究体系建立到完善过程中, 基因组应不会成为限制因素。

4.2 生物信息分析关键点

时珍法的基石是通过生物信息学分析物种基因组, 筛选得到特异靶标序列。为了更准确地筛选物种特异靶标序列, 提出以下三点建议: ① 为减少种内变异对结果的影响, 研究人员在实验技术要求、物种特异两个筛选条件之间可以增加一个不包含种内变异位点的筛选条件。构建物种片段序列库后, 可将库中序列与该物种全部已公开基因组序列进行比对, 仅保留库中能与物种全部公开基因组序列完全匹配的特异靶标序列, 以保证库中所有片段序列的种内保守性。再将库中每一条序列与其他物种的所有片段序列比对, 进行库中片段序列的特异性考察; ② 为方便后续引物设计以及提升方法适用性, 可优先选择在种内保守性高且种间差异性大的区域筛选物种特异靶标序列; ③ 选择与其他物种片段序列相差至少 N 个碱基差异的序列, N 通常大于等于3。可以通过增加差异碱基数量来提高靶标序列的特异性。

4.3 数据库的构建

为简化操作, 时珍法拟建立一个在线数据库网站, 数据库收载所有真核生物物种特异靶标序列, 以供他人参考使用。在数据库构建之前, 需确保获取的物种特异靶标序列和数据库中其他序列准确性, 即保证物

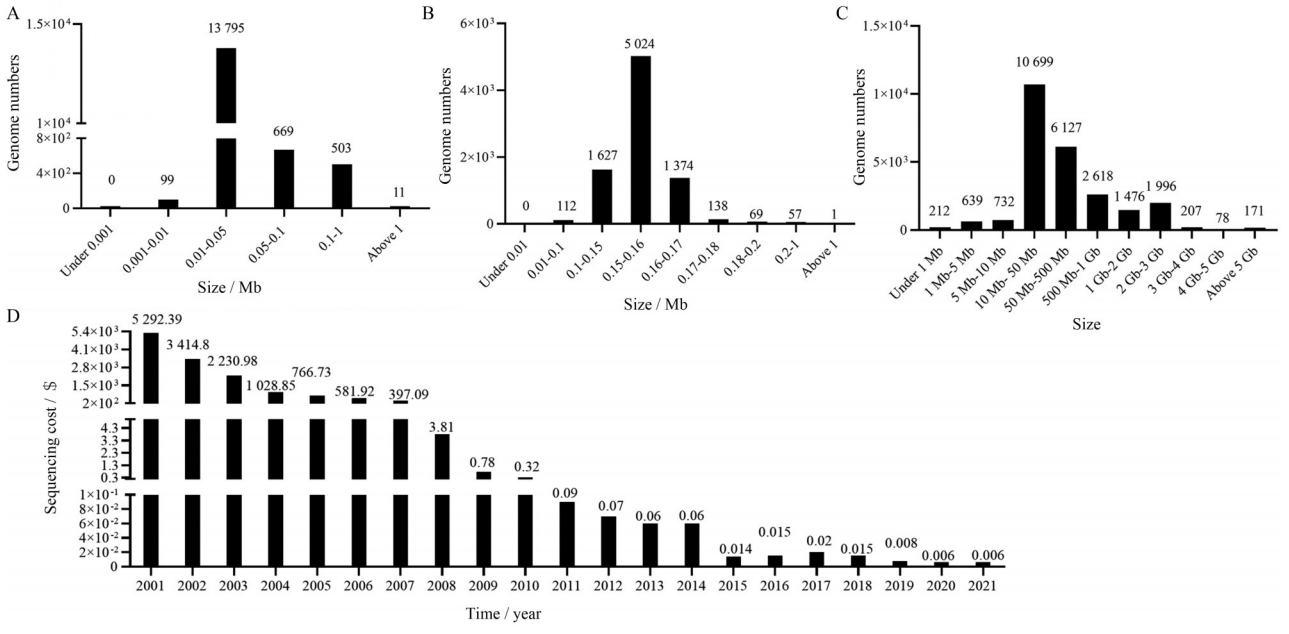

Figure 4 Eukaryotic genome size and sequencing cost statistics. A: Size statistics of eukaryotic mitochondrial genomes according to NCBI; B: Size statistics of eukaryotic chloroplasts genomes according to NCBI; C: Size statistics of eukaryotic nuclear genomes according to NCBI; D: Change in sequencing cost of 1 Mb genomic data from 2001 to 2021

种基因组序列和物种生物样品等材料的正确性: 基因组测序和方法体系构建时使用的生物样品应来自经知名形态鉴定专家鉴定为目标物种的成熟个体或者国内外公认菌种保藏中心的菌株。且物种取样数量应最大限度地覆盖物种整个遗传变异范围。一般来说, 每个物种取3个及以上的个体。完整的数据库构建过程包括三个重要环节: 数据库构建、数据库维护以及数据库使用。

时珍法数据库构建的主要步骤如下: ① 收录所有真核生物物种基因组, 组成物种基因组库; ② 根据不同实验技术对靶标序列的要求, 提取基因组内所有符合条件的片段序列, 得到使用不同实验技术的物种候选片段序列库; ③ 按实验技术分类, 将目标物种所有候选片段序列在其他所有物种的候选片段序列文库中进行比对, 保留物种特异序列组成特异靶标序列库; ④ 将所有特异靶标序列库按物种以及实验技术分类整合, 对其进行处理, 得到靶标序列对应的扩增引物序列和实验操作直接使用的序列; ⑤ 将特异靶标序列及其对应的扩增引物序列、实验操作直接使用的序列和所有实验技术所需的仪器试剂、操作流程以及方法体系一并整合上传到开放的在线网站。因此数据库内容主要包括: 物种基因组库、物种候选片段序列库、物种特异靶标序列库、特异靶标序列对应的扩增引物序列、实验操作直接使用的序列和所有实验技术需要的仪器试剂、操作流程以及方法体系。

数据库维护工作内容主要涉及数据库定时更新。公开基因组越多, 对特异靶标序列的筛选越可靠, 可能会使特异靶标序列发生变化。随着科研人员扩大基因组研究的物种范围, 数据库维护人员需要及时发掘新公开的基因组, 即时更新网站数据库中特异靶标序列及其相关的其他序列。当新的物种基因组公开后, 原有特异靶标序列库应与新公开的基因组进行比对分析, 确保库中序列的物种特异性。数据库维护人员将新基因组收录进物种基因组库, 将其按照建库步骤②处理, 得到新基因组候选片段序列库; 再将之前构建的物种靶标序列库中所有序列在得到的新基因组候选片段序列库比对, 筛选特异序列, 组成更新后的物种特异靶标序列库。最后按建库步骤④、⑤处理得到所有实验技术的物种特异靶标序列库, 即完成数据库维护。更新频率依据基因组公开频率确定, 如每3个月更新一次。

所有人可通过在线网站进入时珍法数据库, 使用和下载其中数据。数据库具体使用方法为: 用户搜索网址进入数据库, 在网站搜索框中输入物种拉丁名, 进行实验操作技术选择后, 即可得到该技术下此物种所有特异靶标序列、实验操作直接使用的序列以及扩增引物序列。同时此网页页面还有该实验技术的仪器试剂、方法流程以及方法体系三个模块。点击每个模块, 即可获取对应内容。通过数据库能获取所有信息, 用户无需再查阅其他资料, 可按其提供的信息直接进行实验。

以物种西红花、实验技术CRISPR-Cas12a作为用户使用数据库网站操作实例, 进行具体阐述 (图5[36])。用户搜索网站进入数据库后, 在搜索框中输入西红花物种拉丁名 *Crocus sativus*, 在下一个实验操作技术选择界面点击CRISPR-Cas12a技术, 即可出现所有与CRISPR-Cas12a技术联用的西红花特异靶标序列、对应的crRNA和引物序列。点击该页面的仪器列表、操作流程以及方法体系模块, 即可获取相应信息。用户根据网站提供的所有信息进行实验操作, 即可检测样品中是否含有西红花。

用户直接从网站下载符合鉴定需求的物种特异靶标序列, 可省去收集和分析基因组等前期基础工作, 方便快捷地进行后续实验操作。实时在线的数据库可避免研究人员进行重复的工作, 浪费大量人力物力。数据库的构建很大程度上简化方法, 减少用户在生物信息分析操作层面耗费的精力和时间, 扩大时珍法的用户群体, 使生物信息知识相对薄弱的用户也能成为方法的潜在使用人群。

该数据库除能够简化时珍法生物信息分析步骤之外, 还具有多个领域潜在应用价值。例如, 海关可从数据库中下载防疫重点关注物种的特异靶标序列, 以选择适合的实验技术检验出入境货物, 防止外来物种入侵。中药生产企业可从数据库中汇总药材基原物种特异靶标序列, 以选择适合的实验技术检验收购药材真伪。公安机关、药品监督管理、检验检疫等机构可对无法鉴定的生物样品进行高通量测序后, 将得到的大量序列在数据库中比对, 实现样品物种鉴定。

5 展望

5.1 AGE具有前沿性和可行性

多项证据表明目前正处于基因组时代: 被誉为"下一个生物学登月计划"的地球生物基因组计划、获得2022年诺贝尔生理学或医学奖的古基因组学等。在这个时代中, 随着测序技术的迅猛进步, 获取物种基因组的难度大大降低, 生物信息技术的发展使研究者能够深入挖掘基因组背后的奥秘。

在COVID-19疫情期间, 为完成国家全员新冠病毒核酸筛查任务, 作为新型冠状病毒核酸检测的必备仪器荧光定量PCR仪在疫情期间被大量采购。疫情结束后该如何处置这些大量闲置且价格不菲的仪器, 时珍法给出解决方案: 以荧光定量PCR仪器实现时珍法来展开鉴定工作。此举可为给因疫情额外大量购入

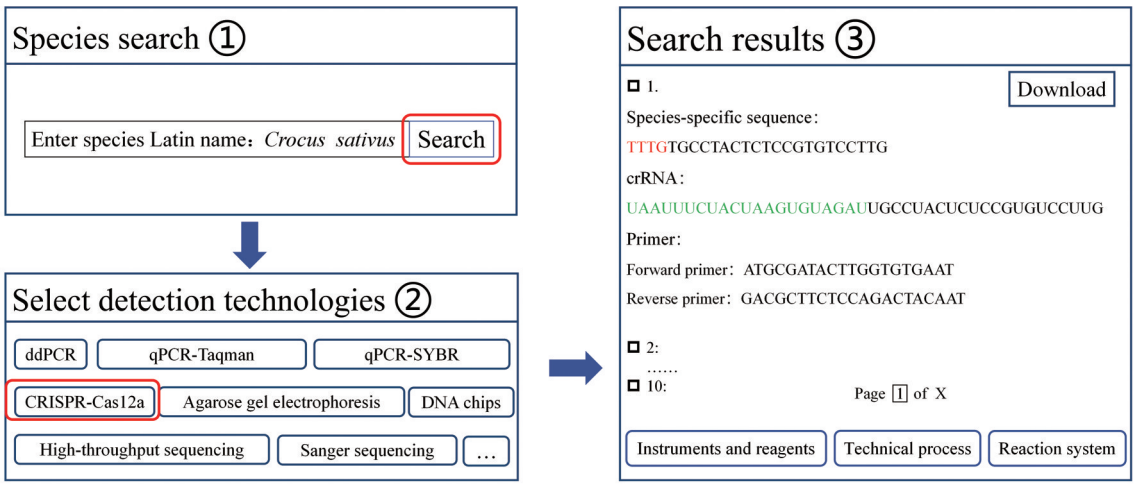

Figure 5 Schematic diagram of users search for saffron-specific target sequences tested using CRISPR-Cas12a at the AGE database website. The species identification is based on this specific target sequence library. Saffron-specific target sequences, crRNA and primer sequences were reprinted with permission from reference[36], copyright (2022) Lijun Hao et al

的荧光定量PCR仪赋予新的活力, 提高仪器的价值, 增加购买产生的超额收益, 避免资金浪费和资源空置, 使得方法有所成, 仪器有所用, 买家有所获, 实现三方共赢的局面。

5.2 生物信息分析软件的开发

如果分析比较基因组的生物信息软件开发成功, 将目标物种和其他物种基因组上传到电脑或服务器中软件进行分析, 调整软件的不同参数, 例如片段序列长度, 是否含特殊结构等, 就可以输出满足用户要求的物种特异靶标序列。该软件具有提取基因组中特定序列和比对两个功能, 可以协助无生物信息分析基础的用户在新基因组公开或自行测得基因组后, 处理时珍法生物信息分析部分工作。用户将新基因组和现有的物种靶标序列库中所有序列输入软件后, 利用软件的两个功能, 按照生物信息分析的实施策略处理, 即可得到更新后的物种特异靶标序列库。只要能够获取物种基因组, 无生物信息学基础的用户都可使用该软件来设计、筛选物种的特异靶标序列。该软件能够扩大时珍法研究体系建立的队伍规模, 降低参与方法研究的门槛。

5.3 商品试剂盒的开发

目前以CRISPR系统为实验技术的时珍法已在植物物种鉴定中得到实现[36]。且CRISPR系统在鉴定领域中可通过可视荧光、试纸条等多种方式获取检测结果。如果将时珍法同成熟的CRISPR系统相结合, 借助其丰富的展现方式, 尤其是试纸条的设计, 今后可以使用商品试剂盒实现物种鉴定。试剂盒里面包含试剂和试纸条, 按照说明书取待检样品的一部分放入试剂中, 再将试纸条蘸取反应后的试剂, 直接读取试纸条的结果即可实现鉴定。商品试剂盒不依赖仪器, 便携性强, 使时珍法应用场景从安置仪器的实验室拓展至户外环境。商品试剂盒的开发可增强时珍法的适用性, 拓宽用户群体。对用户而言, 无需了解试剂盒原理, 直接按照说明书操作即可实现物种鉴定。

5.4 方法的应用前景展望

生物信息软件的开发降低了建立时珍法研究体系的难度, 商品试剂盒的开发降低了方法的受众门槛。两种技术分别从研究和应用两个层面对方法进行简化, 使其具有更广阔的应用前景。鉴于时珍法潜在的高特异性、广适用性等优点, 不受应用对象、使用人员和应用场景等因素限制。

时珍法适用对象囊括植物界、动物界以及真菌界。目前时珍法在植物界已得到应用, 完成不同纲植物间的方法体系构建[36]。在此基础上, 本课题组已成功构建时珍法在动物界和真菌界的物种鉴定体系并已申请2项PCT国际专利、3项中国发明专利, 其中2项PCT专利申请的中国国家阶段申请已获得授权[37,38]。值得一提的是, 传统的鉴定方法和其他分子鉴定方法无法解决的问题, 例如曲霉属和大黄属中某些密切相关物种鉴定, 均通过时珍法成功解决, 并申请专利进行保护。除亲缘关系较远的生物, 时珍法在近缘物种间的区分应也会有良好的表现。时珍法有潜力应用于近缘物种以及种下等级的分子鉴定领域。

时珍法在学科和行业中具有强大的发展潜力。它的出现表明鉴定任务不是只能由具有丰富鉴定知识、熟练操作技能的实验人员负责。该方法作为一种新兴的鉴定方法可能出现在生物科学、食品科学与工程、自然保护与环境生态、中药学、药学、公共卫生与预防医学、公安技术等学科课堂中, 也将在餐饮业、农业、批发

和零售业、卫生和社会工作等行业中得到实践应用。消费者对购入食品的检验、田间生物的快速识别、工厂或者经销商对收购药材的检验、警局对犯罪现场物品勘察鉴别和海关对出入境货物进行检验检疫等场景都可能是时珍法的应用场景。时珍法使鉴定工作不再仅限于科研人员, 实现广泛参与的可能性, 将鉴定的责任和任务扩展至每个人。时珍法将让鉴定走出实验室, 走近各行各业, 走进普通人的生活。

作者贡献: 甘雨桐负责稿件撰写与修改; 辛天怡、许文杰、郝利军、齐桂红和娄千负责稿件修改; 宋经元负责文章整体思路构思和稿件修改等工作。

利益冲突: 所有作者均声明不存在任何利益冲突。

References

[1] Thomson SA, Pyle RL, Ahyong ST, et al. Taxonomy based on science is necessary for global conservation [J]. PLoS Biol, 2018, 16: e2005075.

[2] Bardgett RD, Van Der Putten WH. Belowground biodiversity and ecosystem functioning [J]. Nature, 2014, 515: 505-511.

[3] Dirzo R, Young HS, Galetti M, et al. Defaunation in the Anthropocene [J]. Science, 2014, 345: 401-406.

[4] Lewin HA, Robinson GE, Kress WJ, et al. Earth BioGenome Project: sequencing life for the future of life [J]. Proc Natl Acad Sci U S A, 2018, 115: 4325-4333.

[5] Xin TY, Zhang Y, Pu XD, et al. Trends in herbgenomics [J]. Sci China Life Sci, 2019, 62: 288-308.

[6] Sanger F, Air GM, Barrell BG, et al. Nucleotide sequence of bacteriophage φX174 DNA [J]. Nature, 1977, 265: 687-695.

[7] Lewin HA, Richards S, Lieberman Aiden E, et al. The Earth BioGenome Project 2020: starting the clock [J]. Proc Natl Acad Sci U S A, 2022, 119: e2115635118.

[8] Xu ZC, Li Z, Ren FM, et al. The genome of *Corydalis* reveals the evolution of benzylisoquinoline alkaloid biosynthesis in Ranunculales [J]. Plant J, 2022, 111: 217-230.

[9] Chen JS, Ma E, Harrington LB, et al. CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity [J]. Science, 2018, 360: 436-439.

[10] Li SY, Cheng QX, Wang JM, et al. CRISPR-Cas12a-assisted nucleic acid detection [J]. Cell Discov, 2018, 4: 20.

[11] Xu WJ, Zhu PY, Xin TY, et al. Droplet digital PCR for the identification of plant-derived adulterants in highly processed products [J]. Phytomedicine, 2022, 105: 154376.

[12] Parlapani FF, Syropoulou F, Tsiartsafis A, et al. HRM analysis as a tool to facilitate identification of bacteria from mussels during storage at 4 ℃ [J]. Food Microbiol, 2020, 85: 103304.

[13] Lou Q, Xin TY, Xu WJ, et al. TaqMan probe-based quantitative real-time PCR to detect *Panax notoginseng* in traditional Chinese patent medicines [J]. Front Pharmacol, 2022, 13: 828948.

[14] Chen YH, Liu LY, Tsai WH, et al. Using DNA chips for identification of tephritid pest species [J]. Pest Manag Sci, 2014, 70: 1254-1261.

[15] Hebert PDN, Cywinska A, Ball SL, et al. Biological identifications through DNA barcodes [J]. Proc Biol Sci, 2003, 270: 313-321.

[16] Yang B, Zhang ZX, Yang CQ, et al. Identification of species by combining molecular and morphological data using convolutional neural networks [J]. Syst Biol, 2022, 71: 690-705.

[17] Raick X, Huby A, Kurchevski G, et al. Use of bioacoustics in species identification: Piranhas from genus *Pygocentrus* (Teleostei: Serrasalmidae) as a case study [J]. PLoS One, 2020, 15: e0241316.

[18] Edwards K, Manley M, Hoffman LC, et al. Differentiation of South African game meat using near-infrared (NIR) spectroscopy and hierarchical modelling [J]. Molecules, 2020, 25: 1845.

[19] Wang HQ, Song W, Tao WW, et al. Identification wild and cultivated licorice by multidimensional analysis [J]. Food Chem, 2021, 339: 128111.

[20] Ho CS, Jean N, Hogan CA, et al. Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning [J]. Nat Commun, 2019, 10: 4927.

[21] Ahmed FE. Detection of genetically modified organisms in foods [J]. Trends Biotechnol, 2002, 20: 215-223.

[22] Razin S, Rottem S. Identification of *Mycoplasma* and other microorganisms by polyacrylamide-gel electrophoresis of cell proteins [J]. J Bacteriol, 1967, 94: 1807-1810.

[23] Wolf C, Rentsch J, Hübner P. PCR−RFLP analysis of mitochondrial DNA: a reliable method for species identification [J]. J Agric Food Chem, 1999, 47: 1350-1355.

[24] Mueller UG, Wolfenbarger LL. AFLP genotyping and fingerprinting [J]. Trends Ecol Evol, 1999, 14: 389-394.

[25] Cottenet G, Blancpain C, Sonnard V, et al. Two FAST multiplex real-time PCR reactions to assess the presence of genetically modified organisms in food [J]. Food Chem, 2019, 274: 760-765.

[26] Savolainen V, Cowan RS, Vogler AP, et al. Towards writing the encyclopaedia of life: an introduction to DNA barcoding [J]. Philos Trans R Soc Lond B Biol Sci, 2005, 360: 1805-1811.

[27] Xin TY, Li XW, Yao H, et al. A two-dimensional DNA barcode system for circulation regulation of traditional Chinese medicine [J]. Sci Sin Vit (中国科学: 生命科学), 2015, 45: 695-702.

[28] Chen SL, Yin XM, Han JP, et al. DNA barcoding in herbal medicine: retrospective and prospective [J]. J Pharm Anal, 2023, 13: 431-441.

[29] Lou Q, Xin TY, Song JY. Application of DNA barcoding technology in the whole industrial chain of traditional Chinese medicine [J]. Acta Pharm Sin (药学学报), 2020, 55: 1784-1791.

[30] Li RJ, Wu LW, Xin TY, et al. Analysis of chloroplast genomes and development of specific DNA barcodes for identifying the original species of Rhei Radix et Rhizoma [J]. Acta Pharm Sin

(药学学报), 2022, 57: 1495-1505.

[31] Xiang L, Tang H, Cheng JL, et al. The species traceability of the ultrafine powder and the cell wall-broken powder of herbal medicine based on DNA barcoding [J]. Acta Pharm Sin (药学学报), 2015, 50: 1660-1667.

[32] Mora C, Tittensor DP, Adl S, et al. How many species are there on earth and in the ocean [J]. PLoS Biol, 2011, 9: e1001127.

[33] Nevill PG, Zhong X, Tonti-Filippini J, et al. Large scale genome skimming from herbarium material for accurate plant identification and phylogenomics [J]. Plant Methods, 2020, 16: 1.

[34] Liu HL, Wang XB, Wang GB, et al. The nearly complete genome of *Ginkgo biloba* illuminates gymnosperm evolution [J]. Nat Plants, 2021, 7: 748-756.

[35] Pellicer J, Fay MF, Leitch IJ. The largest eukaryotic genome of them all [J]. Bot J Linn Soc, 2010, 164: 10-15.

[36] Hao LJ, Xu WJ, Qi GH, et al. GAGE is a method for identification of plant species based on whole genome analysis and genome editing [J]. Commun Biol, 2022, 5: 947.

[37] Song JY, Hao LJ, Xu WJ, et al. Method and use for identifying plant species based on whole genome analysis and genome editing: CN, CN115843318A [P]. 2023-03-24.

[38] Song JY, Qi GH, Xu WJ, et al. Method and use for identifying eukaryotic species based on whole genome analysis: CN, CN115087750A [P]. 2023-05-02.