



余岩,四川大学生命科学学院教授,博士生导师,四川大学2012届博士。教育部重要人才计划青年项目入选者,四川省学术和技术带头人后备人选,获得四川省青年科学基金A类项目资助。研究方向为植物系统发育与演化,致力于演化理论与方法学研究,提出了物种起源扩散、基因组数据挖掘及细胞学分析的理论框架与工具集,成果应用于物种演化模式、基因功能解析和物种数据库建设等领域。作为第一作者或通信作者在MBE、MER、MPE等进化生物学顶级期刊发表SCI论文30余篇,被引5000余次。主持国家自然科学基金面上项目3项、国家自然科学基金青年项目、国家基础条件平台等各级项目十余项,出版教材和专著5部,获得教育部自然科学二等奖、四川大学教学成果一等奖、“好未来”教学名师等荣誉。

Phylox:植物二三代测序数据的DNA条形码 自动化鉴定系统

刘振宇^{1,2}, 郑羽皓¹, 宋永秀³, 余岩^{1,2}

- (1. 四川大学生命科学学院生物资源与生态环境教育部重点实验室, 西南特色生物资源研究与利用四川省重点实验室, 山地植物迁地保护与资源利用实验室, 成都610064;
2. 成都市植物园, 成都市公园城市植物科学研究院, 成都610083;
3. 中国科学院武汉植物园植物多样性与特色经济作物国家重点实验室, 武汉430074)

摘要: 针对传统植物分子鉴定在处理复杂类群及海量高通量测序数据时存在的碎片化组装与自动化程度低等难题,本研究开发了集成二代与三代测序数据的植物DNA条形码自动化提取与鉴定软件Phylox。该软件以统一的本地化Rust计算引擎为核心,构建了覆盖被子植物216个科的本地分层参考数据库,并提供跨平台图形用户界面(GUI)。对于二代Illumina数据,软件采用GeneMiner 3同构的加权De Bruijn图靶向恢复流程,支持标准、单拷贝与倍性感知三种组装模式,以及全基因组、捕获测序与转录组三种数据类型,有效抑制旁系同源导致的误组装;对于三代Nanopore长读段数据,采用UNOISE3去噪直接输出真实观测序列(zero-radius OTU, zOTU),结合属级一致性校验剔除污染与嵌合序列,并通过ITS与*psbA-trnH*差异位点加权的多基因联合鉴定提升近缘类群鉴定的鲁棒性。在伞形科28个近缘种样本测试中,系统实现了100%的属级与92.86%的种级分类准确率;对单标记及混合标记三代样本的测试进一步验证了多标记分离鉴定的可靠性。在基准测试中,基于金标准已知的模拟二代扩增子与模拟三代ONT数据集,Phylox在10个样本-标记对上全部恢复目标条形码。该软件实现了从原始测序数据输入至物种鉴定报告输出的一键式闭环处理,其安装指南、详细文档及源代码均已在Gitee开源(<https://gitee.com/sculab/phylox>),为植物资源调查、生物多样性监测及复杂类群的分子鉴定提供了具有工程化应用价值的工具。

关键词: DNA条形码; 高通量测序; 分子鉴定; ITS; *psbA-trnH*

中图分类号: Q811.4 **文献标志码:** A **DOI:** 10.19907/j.0490-6756.260183

收稿日期: 2026-05-31 **修回日期:** 2026-08-12 **接受日期:** 2026-08-12

基金项目: 科学技术基础资源调查计划(2022FY101000); 中国国家自然科学基金(32271552、32571738); 四川自然科学基金会(2026NSFSCZY0116); 中央高校基本科研业务费(CJ202602)

作者简介: 刘振宇(2004—), 男, 安徽六安人, 硕士研究生, 主要研究方向为植物分子鉴定. E-mail: 18155295565@163.com

通信作者: 余岩. E-mail: yyu@scu.edu.cn

Phylox: An automated system for identifying plant DNA barcodes from Second- and Third-Generation sequencing data

LIU Zhenyu^{1,2}, ZHENG Yuhao¹, SONG Yongxiu³, YU Yan^{1,2}

(1. Key Laboratory of Bio-Resources and Eco-Environment of Ministry of Education; Southwest Bio-resources R&D Key Laboratory of Sichuan Province; Joint Laboratory for Ex Situ Conservation and Resource Utilization of Mountain Plants, College of Life Sciences, Sichuan University, Chengdu 610064, China;

2. Chengdu Botanical Garden, Chengdu Park City Botanical Science Research Institute, Chengdu 610083, China;

3. State Key Laboratory of Plant Diversity and Prominent Crops, Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan 430074, China)

Abstract: To address fragmented assembly and limited automation in conventional plant molecular identification when complex taxa and large high-throughput sequencing datasets are processed, Phylox was developed as automated software for plant DNA barcode extraction and identification using both second- and third-generation sequencing data. A unified local Rust computing engine was used, a local hierarchical reference database covering 216 angiosperm families was built, and a cross-platform graphical user interface (GUI) was provided. For second-generation Illumina data, a weighted De Bruijn graph-based targeted recovery workflow consistent with GeneMiner 3 was used. Three assembly modes (standard, single-copy, and ploidy-aware) and three data types (whole-genome, target-capture, and transcriptome data) were supported, reducing misassembly caused by paralogs. For third-generation Nanopore long-read data, UNOISE3 denoising was used to directly produce observed sequences as zero-radius OTUs (zOTUs). Contaminant and chimeric sequences were removed through genus-level consistency checks, and closely related taxa were identified by multi-gene analysis weighted according to variable sites in ITS and *psbA-trnH*. In tests of 28 samples from closely related Apiaceae species, genus-level and species-level classification accuracies of 100% and 92.86%, respectively, were obtained. The reliability of multi-marker separation and identification was further confirmed using third-generation single-marker and mixed-marker samples. In benchmarks using simulated second-generation amplicon and third-generation ONT datasets with known ground truth, target barcodes were recovered for all 10 sample-marker pairs. A one-click workflow from raw sequencing data to a species identification report was implemented. The installation guide, detailed documentation, and source code were released on Gitee (<https://gitee.com/sculab/phylox>). The software can be used for plant resource surveys, biodiversity monitoring, and molecular identification of complex taxa.

Keywords: DNA barcode; high-throughput sequencing; molecular identification; ITS; *psbA-trnH*

植物分子鉴定是植物资源开发、生物多样性保护及生态系统监测的核心技术支撑。传统的形态学鉴定方法易受表型可塑性、隐存种及样本退化等因素制约,难以满足大规模、高通量背景下的精准鉴定需求^[1-2]。自DNA条形码技术提出以来,利用标准化、通用性强的短DNA片段实现物种的快速、准确鉴定,已成为植物分类学、生态学及中药材基原鉴定的关键手段^[3-6]。其中,核糖体DNA内部转录间隔区ITS因高拷贝数、高进化速率及显著的种间分辨率,在被子植物鉴定中应用最为广泛;叶绿体*psbA-trnH*间隔区则常作为互补条形码

码,在近缘种的精细区分中发挥重要作用^[7-11]。

随着高通量测序技术的兴起,二代测序平台凭借高通量与低成本优势,成为植物DNA条形码研究的主流,为大规模扩增子测序和群体遗传分析提供了海量短读段数据^[12-13];其中,基于基因组浅层测序(genome skimming)的条形码获取策略与既有条形码体系向后兼容,且可随测序成本下降不断加深覆盖,已成为高通量获取条形码序列的重要途径^[14]。然而,受限于读长较短,二代测序在跨越基因组重复区域和重建完整条形码时易出现组装碎片化,且难以有效克服复杂类群测序中的

扩增与组装偏好性难题,尤其在处理近缘种复合体时,分辨率不足限制了目标条形码的自动化提取效率。近年来,以长读长为核心优势的三代测序技术有效弥补了上述缺陷,其能够实现条形码区域的单分子全长覆盖,显著降低组装偏差,并提升混合样本中目标序列的分离度^[15-19]。尽管三代测序在单分子准确率及数据处理复杂度上仍存在挑战,但其与二代测序在海量高通量数据与长读长全长重建方面的互补,正推动植物分子鉴定迈向高分辨率的新阶段^[20]。

随着测序数据的爆发式增长,国内外针对二代测序数据的植物DNA条形码提取与鉴定工具已取得显著进展。在序列比对层面,经典工具BLAST仍被广泛应用;同时,深度学习技术在系统发育信号挖掘中展现出巨大潜力,如基于预训练DNA语言模型的PhyloTune实现了系统发育树的快速更新,而基于k-mer频率表征的varKoder则利用Vision Transformer模型实现了二代原始数据的直接鉴定^[21-22]。针对特定平台,虽然基于二代测序的QIIME 2、AmpliPiper等扩增子分析流程已趋于成熟,但目前专用于植物长读段条形码数据自动化鉴定的分析工具仍较匮乏,缺乏一体化 workflow^[23-24]。因此,摆脱对传统从头组装流水线的依赖,开发一套能够协同处理二三代原始异质数据、兼顾计算效率与分辨率的全流程自动化分析平台,是该领域亟须解决的关键问题^[25-26]。本团队前期开发的GeneMiner工具能够从低覆盖度二代测序数据中高效挖掘目标标记基因,为下游鉴定流程提供了数据基础^[27];其新一代内核GeneMiner 3进一步提供了多组装模式与倍性感知(phased)的靶向基因恢复能力,为本研究二代模块的算法升级奠定了基础。

针对上述问题,本研究开发了Phylox软件,其集成二代与三代测序数据,实现ITS及*psbA-trnH*自动提取与物种鉴定。鉴于植物DNA条形码参考数据库长期存在构建零散、缺乏系统质控的问题^[28],且参考数据的自动化管理与维护直接关系鉴定可靠性^[29],为保证鉴定精度和适用性,我们首先构建了高质量的本地参考序列数据库,并在此基础上开发了针对二代数据的科级预筛选-精细提取一体化算法,以及针对三代长读长数据的ITS/*psbA-trnH*自动分离、去噪优化与zOTU序列生成流程,实现了从原始测序数据输入到BLAST物种

鉴定结果输出的一键式自动化闭环处理。在基准测试中,本研究使用模拟数据集与多套真实测序数据,将Phylox与varKoder、Phyluce、PhyProbe等工具在同一数据上进行系统对比,给出定量的性能差异。研究为植物分子鉴定提供了从分散工具向集成化工作流转变的解决方案,有助于提升植物基原鉴定、资源调查和生物多样性监测的效率,也为分子系统学向自动化数据分析转变提供了方法参考。

1 材料与方法

1.1 本地参考序列数据库的构建

为实现二代与三代测序数据的高效物种鉴定,本研究构建了覆盖被子植物216个科的本地ITS及*psbA-trnH*参考序列数据库。该数据库采用分级结构设计,旨在平衡鉴定覆盖度与分类特异性。

1.1.1 科级主库构建原则与质量控制

科级数据库涵盖216个被子植物科,每条候选序列均包含ITS1、5.8S rRNA及ITS2构成的完整ITS区域。原始序列通过关键词检索与分类学拓扑过滤自NCBI Nucleotide数据库获取。序列筛选标准如下:序列长度限制在500~1000 bp;剔除含高比例N碱基或连续低质量区域的序列;基于18S和28S rRNA保守边界对序列进行精确截断,去除侧翼非目标区。每个科均筛选具有系统发育代表性的属种序列,以覆盖该科的主要进化分支。构建完成的科级主库用于二代测序数据的科级预筛选及GeneMiner 3目标基因挖掘。

1.1.2 属种级精细子库构建策略

在科级主库基础上,本研究构建了各科的属种级子库。构建策略旨在兼顾系统发育多样性与序列冗余控制:每个物种保留1条代表序列,每个属的序列总数不超过15条;为保证统计分析的稳健性,每个子库至少保留5条有效序列。数据源自NCBI Nucleotide数据库。为适应不同PCR策略及退化样本,子库同时收录了部分物种的独立ITS1、ITS2片段及完整ITS序列。所有序列均经18S和28S rRNA边界修剪,并经人工校核与分类学检查,剔除嵌合体及注释错误序列。该策略在保证分类覆盖度的同时,有效控制了数据库规模,降低了计算负荷。

1.1.3 数据库格式、标准化注释与检索管理

所有参考序列以标准 FASTA 格式存储。在完成基础的物理质控与去冗余后,针对公共数据库中广泛存在的物种分类错误注释与鉴定失误序列,进一步构建了基于系统发育拓扑冲突的本地化数据动态纯化算法。该算法依据分类学一致性机理,针对多级分层数据库中的每个属级分类单元,调用 vsearch 构建属内全对齐距离矩阵并快速组装局部邻接系统发育树。系统通过计算各末端分支的平均演化距离与速率,引入自适应标准差阈值模型对异常拓扑结构进行动态扫描。当某条参考序列的属内遗传距离显著偏离总体均值,且其最邻近节点在拓扑结构上归属于异源科属时,系统将其判定为高风险错误注释并执行自动化剪枝与剔除。序列标题经过统一规范化处理,包含物种名称、完整分类阶元、GenBank 登录号和序列类型等信息。数据库采用分级目录管理,科级主库置于根目录,各科子库置于对应子文件夹,便于程序索引与调用。同时,利用 makeblastdb 工具为所有序列集建立本地 BLAST 索引,以支持并行比对。

1.1.4 三代测序专用数据库及竞争性比对索引构建

为适应三代长读段数据的鉴定与目标序列分离,本研究构建了 ITS 和 *psbA-trnH* 专用本地数据库。ITS 区域采用 NCBI 官方预格式化的真核生物 ITS 参考数据集。*psbA-trnH* 则通过 NCBI Nucleotide 数据库进行定长及关键词检索,筛选获得约 14 万条植物参考序列。为实现基于竞争性比对的目标条形码分离,对两套数据库进行处理:为 ITS 和 *psbA-trnH* 序列标识符分别添加唯一前缀以避免 ID 冲突,合并序列集,并利用 minimap2 软件构建竞争性比对索引,用于长读段数据的目标序列分离。

1.2 软件架构设计与自动化分析流程

Phyllox 以统一的本地化 Rust 计算引擎为核心,并辅以基于 Tauri 的跨平台图形用户界面。顶层调度器根据读段类型与用户配置,自动将二代 Illumina 数据路由至靶向基因恢复流水线、将三代 Nanopore 长读段路由至条形码分离与鉴定流水线,实现从原始数据到物种鉴定报告的全自动化处理(图 1)。

1.2.1 二代测序数据 ITS 自动化分层提取算法

二代测序自动化流程主要针对被子植物 Illumina 平台产生的双端短读段数据设计,其底层基于本团队 GeneMiner 3 同构的加权 De Bruijn 图(wDBG)内核,提供标准(standard)、单拷贝(single_copy)与倍性感知(phased)三种组装模式,并适配全基因组浅层测序、探针捕获与转录组三种数据类型。该内核通过种子自发现与贪婪扩展实现低覆盖度数据中目标分子标记的精准捕获;针对旁系同源与多拷贝风险,single_copy 模式集成 UCE 多拷贝守卫^[30],依据深度突刺、核心-侧翼深度断崖与覆盖度变异三类判据抑制旁系同源误组装;phased 模式则通过 minimap2 比对、Clair3-rs(纯 Rust ONNX 变异检测)或 freebayes 变异识别、whatsaph 定相与 bcftools 一致性重构,按估计倍性 1~6 输出单倍型序列。鉴于提取效能对参考序列与靶标样本间的系统发育亲缘关系高度敏感,本研究设计了由粗及细的双级迭代提取方案。

在初筛阶段,软件调用覆盖 216 个被子植物科的科级主库作为广谱参考模板,以提升未知分类地位样本的初始捕获率。初步组装的 ITS 候选序列通过本地 BLAST 算法与科级主库比对,并综合序列一致性、覆盖度、E-value 及 bitscore 等指标进行加权评估,锁定最佳匹配科级阶元;为降低传统比对方法的错配风险,原始双端数据可同步交由深度学习鉴定软件 varKoder 处理,利用其基于 k-mer 频率特征的系统发育分类结果对 BLAST 结果进行交叉验证。在精细提取阶段,系统依据初筛所得分类地位自动选择组装模式:对拷贝数敏感的物种调用单拷贝模式,对多倍体或倍性不明的样本调用多拷贝模式,随后加载对应的属种级子库进行高特异性二次提取与组装,最终输出规范化命名的 ITS 序列及相关特征元数据。流程为每个样本配置独立工作空间并输出标准化日志,以确保分析过程的可追溯性与可复现性。

1.2.2 三代测序长读段数据自动化分离与优化流程

三代测序自动化流程基于长读段物理特征设计,旨在解决混合原始 FASTQ 文件中 ITS 与 *psbA-trnH* 条形码标记的自动分流与去噪序列重建问题。流程起始端利用 minimap2 将长读段比对至合并参考序列索引,通过竞争性比对机制实现分流。调用 seqkit 工具进行正负链取向校正,使用

cutadapt工具切除poly-A/T尾部,并支持根据实验配置动态剔除引物序列,同时结合长度阈值过滤低质量片段,获取高纯度清洁读段集。

针对三代长读长测序数据高频发生单碱基插入、缺失及随机替换错误的固有缺陷,本系统采用UNOISE3去噪算法^[31-32]处理分流后的异质长读段。UNOISE3基于丰度排序与差异距离阈值去噪原理,通过vsearch的--cluster_unoise算子对清洁读段执行去噪聚类,直接输出高置信度的真实观测序列(zero-radius OTU, zOTU),而非传统聚类后的共识(consensus)序列。其核心假设在于:真实生物序列在样本中呈现高丰度富集,而测序错误以低频随机模式分布,可通过丰度阈值与差异距离有效分离。与共识重构策略相比,zOTU方法保留了每一高丰度簇的原始观测读段,避免了共识投票过程可能引入的序列重组伪影,同时显著降低计算复杂度。随后,系统提取丰度高于设定阈值(默认 ≥ 5 条读段)的zOTU作为候选条形码序列。UNOISE3的判别准则同时依赖序列距离与丰度比:对唯一序列 u 与某zOTU中心 c ,设距离为 $d(u, c)$ (不计gap的错配数),则 u 并入 c 的充要条件为 $A(u)/A(c) \leq \beta(d) = 2^{(-1-\alpha \cdot d)}$,其中 $A(u)$ 、 $A(c)$ 分别为二者的丰度, α 为丰度-距离系数(默认2.0)。该式表明离中心越远的序列允许进入簇的丰度上限呈指数下降,而“更近但丰度更高”的序列自成簇,从而在ONT高错误率下把真实生物序列与测序错误分离。

在此基础上,系统进行属级一致性校验以剔除污染与嵌合序列:按各zOTU的丰度加权统计其BLAST最佳命中的属级归属,若某zOTU的最佳属与主流属跨科不一致,则判为污染或嵌合并剔除。最终获取的zOTU序列用于后续鉴定,同时输出包含序列长度、支持读段数与丰度等信息的质控报告。整个流程采用模块化目录管理,并生成详细日志,记录分离效率与聚类信息。

1.2.3 跨平台测序数据的本地与远程BLAST自动化鉴定

在获取目标条形码序列后,系统进入通用的系统发育分类学赋值模块。该模块作为两类测序管道的共享下游,对ITS与 $psbA-trnH$ 分别进行本地BLAST+鉴定后,采用多基因联合物种赋值:当两基因种级鉴定一致时,得分取两基因支持率之积;属级一致时得分乘以0.5;两基因命中集合

无交集时得分乘以0.05;当属内种级结论冲突时给出杂交提示。此外,系统可接入植物图像识别与Species2000中文名数据库,为鉴定结论提供辅助证据与中文名富化。

为确保鉴定结论的科学性,系统针对不同测序平台设计了差异化的质控机制:针对二代短读段数据,系统采用BLAST比对和varKoder深度学习模型鉴定的双轨交叉验证体系,以保障鉴定的准确性;针对三代长读段数据,因其能够直接覆盖条形码全长区域,具备更强的抗错配能力,系统直接采用高信度BLAST比对策略即可获得准确鉴定结果。此外,系统具备高度的灵活性,用户不仅可调用内置标准库,还可指定自定义本地参考集,软件提供了参考序列数据库替换接口。用户能够指定自定义分子标记的FASTA文件路径,自动重新构建本地BLAST索引和minimap2竞争性比对索引,实现对任意目标新标记的一键式自动化捕获与物种赋值。集成了向NCBI nt数据库执行远程并发请求功能,从而为本地数据库覆盖范围之外的罕见或新发植物类群鉴定提供国际化资源支撑。针对非模式物种提取的场景,Phylox提供了多项专门适配:1)多分辨率参考检索,科级主库与属种级子库两级检索,使提取仅依赖该科是否有代表性参考,而非该物种是否已有近缘参考;2)wDBG参考位置加权允许适度分歧,参考仅用于引导种子选择而不强制序列一致性;3)三代低参考依赖,竞争性分离仅需ITS/ $psbA-trnH$ 索引即可工作,无需完整基因组;4)可扩展标记接口,支持自定义标记FASTA自动重建索引。这些设计使Phylox在缺乏近缘参考的非模式物种中仍能稳定工作。

1.2.4 测试数据集的构建

本研究选取28个伞形科(Apiaceae)植物样本构建独立二代测序数据集^[27]。为规避分类学鉴定中的循环论证风险并确保基准数据的可靠性,该测试集样本均源自已公开发表且经过多次分类学互证的经典数据集。所有测试样本的原始测序数据在输入本系统前均剔除预设的物种标签与分类元数据,直接作为分析流的起始输入,以此客观评估本系统对系统发育关系密切的近缘类群的分类解析效能。为进一步定量评估提取与鉴定精度,本研究构建了模拟数据集作为金标准:从本地参考库中选取5个样本(南天竹 *Nandina domestica*、

黄荆 *Vitex negundo*、钮子瓜 *Zehneria scabra*、杏 *Prunus armeniaca*、豆梨 *Pyrus calleryana*), 提取其 ITS 与 psbA-trnH 参考序列作为真实序列; 二代模拟数据使用 wgsim 依据扩增子序列生成 150 bp 双端 reads (覆盖度约 100×、替换错误率 0.5%), 三

代模拟数据按真实 ONT 错误特征生成, 在含引物扩增子序列上引入约 8% 的 homopolymer 集中型错误 (替换与插入/缺失), 每个标记生成 300 条长读段。所有模拟数据均保留样本真实物种标签, 用于后续恢复率、序列一致性与错误率的定量评估。

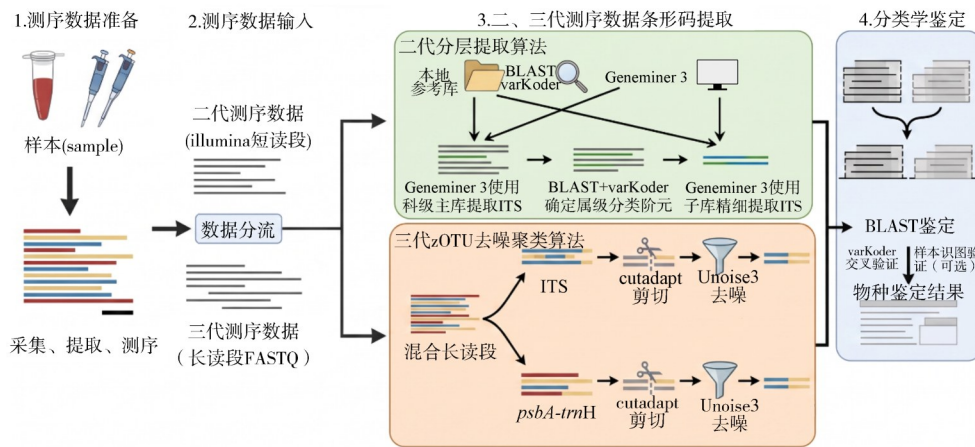
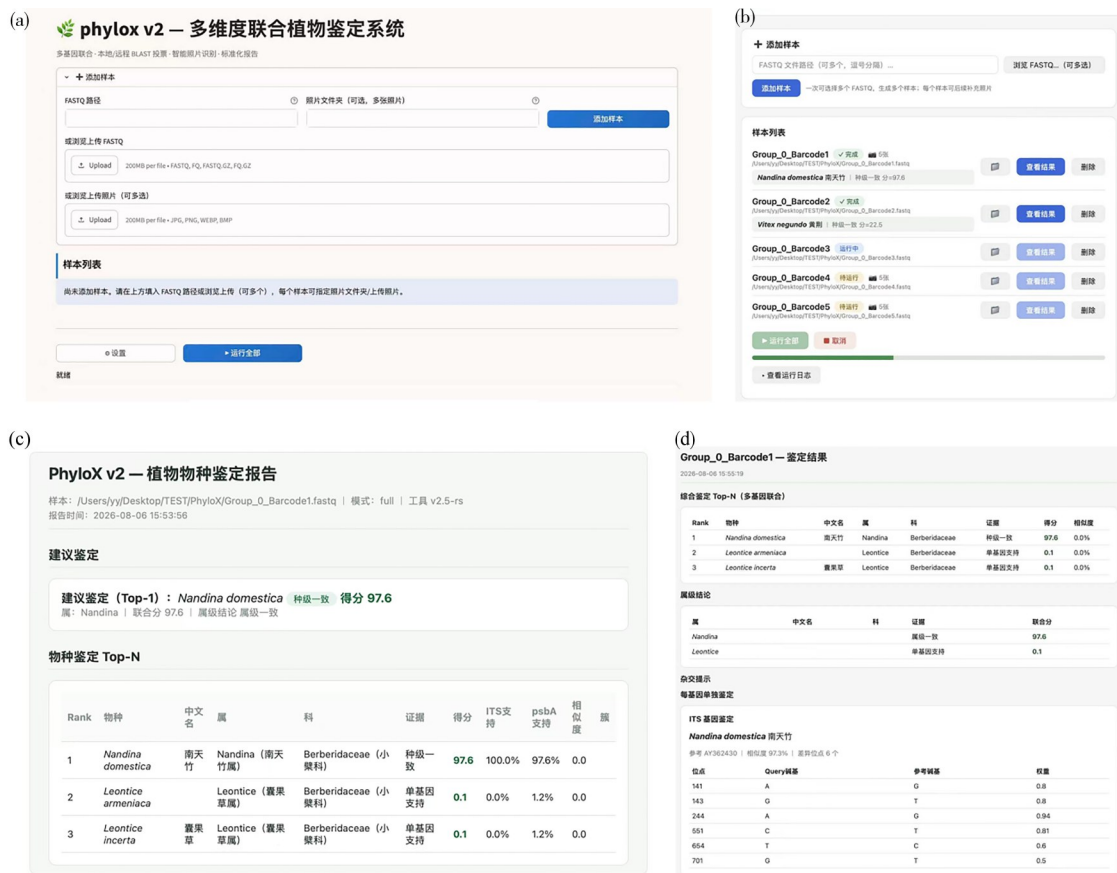


图 1 Phylox 分析流程

Fig. 1 Phylox analysis workflow



(a) 主工作流程界面 (支持 FASTQ 路径选择、参数配置与一键运行); (b) 样本批量处理界面 (多条条形码并行任务管理); (c) Top-N 物种联合鉴定报告 (含中文名、属级结论与综合得分); (d) 每基因差异位点报告 (支持 ITS 与 psbA-trnH 的独立质控)。

图 2 Phylox 软件图形用户界面

Fig. 2 Graphical user interface of Phylox

2 结果与分析

2.1 二代测序 ITS 序列提取与物种鉴定结果

为客观评估本系统在近缘类群中的分类解析效能,本研究在科、属、种 3 个分类阶元上对 28 个二代测试样本的鉴定结果进行统计,同时引入基于 Vision Transformer 与 k-mer 频率变异特征的深度学习物种鉴定框架 varKoder 作为第三方对比工具,结果见表 1;表中给出 Phylox 鉴定得分前 3 名物种以及这些物种在比对得分前 100 位出现的次数,varKoder 的对比结论详见 2.3 节。

实验结果表明,本系统在伞形科这一分类学难点类群中展现出优异的精细分辨率。在 28 个测试样本中,系统实现了 26 个样本的精准种级分类,

准确率达 92.86%;其余 2 个样本虽因种间序列高度保守未能实现种级区分,但均被准确划归至所属属级支系,属级鉴定准确率为 100%。相比之下,varKoder 在精细分类阶元的分辨率随分类阶元细化显著下降,其种级与属级准确率分别仅为 28.57%(8/28)和 64.29%(18/28)。上述对比表明,相比基于全基因组宏观特征表征的深度学习模型,本系统采用的标记基因靶向深度挖掘结合精细子库迭代比对策略,在处理复杂植物类群的精细物种鉴定方面具备显著技术优势。此外,varKoder 在科级水平达到的 100% 准确率,从实验层面验证了将其作为本系统科级预筛选交叉验证模块的可行性与可靠性。

表 1 二代测试集鉴定结果
Tab. 1 Identification results of the second-generation sequencing test dataset

样本编号	形态鉴定结果	Phylox 鉴定结果	varKoder 鉴定结果
1	<i>Aegopodium podagraria</i>	<i>Aegopodium podagraria</i> (0.997 9, 1)	<i>Aegopodium podagraria</i>
		<i>Aegopodium alpestre</i> (0.978 7, 3)	
		<i>Conioselinum schugnanicum</i> (0.972 1, 1)	
2	<i>Angelica acutiloba</i>	<i>Angelica acutiloba</i> (1.0, 7)	<i>Angelica</i>
		<i>Angelica stenoloba</i> (0.998 5, 1)	
		<i>Peucedanum morisonii</i> (0.988 0, 1)	
3	<i>Angelica decursiva</i>	<i>Angelica decursiva</i> (1.0, 2)	<i>Angelica</i>
		<i>Heracleum hemsleyanum</i> (1.0, 1)	
		<i>Angelica gigas</i> (0.983 7, 1)	
4	<i>Anthriscus sylvestris</i>	<i>Anthriscus sylvestris</i> (0.998 1, 65)	<i>Anthriscus sylvestris</i>
		<i>Ferula licentiana</i> (0.996 3, 1)	
		<i>Torilis japonica</i> (0.995 2, 2)	
5	<i>Bupleurum chinense</i>	<i>Bupleurum chinense</i> (1.0, 26)	<i>Bupleurum chinense</i>
		<i>Bupleurum yinchowense</i> (0.997 1, 4)	
		<i>Bupleurum scorzoniferifolium</i> (0.996 5, 3)	
6	<i>Chamaesium paradoxum</i>	<i>Chamaesium paradoxum</i> (1.0, 81)	Apiaceae
		<i>Chamaesium jiulongense</i> (0.990 1, 19)	
		<i>Cnidium monnieri</i> (0.995 7, 1)	
7	<i>Cnidium monnieri</i>	<i>Pimpinella</i> sp. YX-2018 (0.991 7, 1)	Apiaceae
		<i>Kitagawia terebinthacea</i> (0.981 6, 1)	
		<i>Coriandrum sativum</i> (0.998 9, 6)	
8	<i>Coriandrum sativum</i>	<i>Oreocomopsis aromatica</i> (0.951 7, 2)	Apiaceae
		<i>Conioselinum tataricum</i> (0.949 6, 2)	
		<i>Cryptotaenia japonica</i> (1.0, 12)	
9	<i>Cryptotaenia japonica</i>	<i>Sanicula serrata</i> (0.990 9, 1)	Apiaceae
		<i>Cryptotaenia canadensis</i> (0.973 4, 3)	
		<i>Cyclospermum leptophyllum</i> (1.0, 4)	
10	<i>Cyclospermum leptophyllum</i>	<i>Trachyspermum ammi</i> (0.926 5, 1)	Apiaceae
		<i>Seselopsis tianschanica</i> (0.904 4, 2)	
		<i>Daucus carota</i> (1.0, 43)	
11	<i>Daucus carota</i>	<i>Daucus littoralis</i> (1.0, 2)	<i>Daucus carota</i>
		<i>Daucus virgatus</i> (1.0, 2)	
		<i>Anethum foeniculum</i> (1.0, 9)	
12	<i>Anethum foeniculum</i>	<i>Apiaceae</i> sp. EM372 (0.995 6, 2)	<i>Anethum graveolens</i>
		<i>Anethum graveolens</i> (0.983 2, 1)	

(续表 1)			
样本编号	形态鉴定结果	Phylox 鉴定结果	varKoder 鉴定结果
13	<i>Hansenia phaea</i>	<i>Hansenia phaea</i> (1.0, 3)	<i>Hansenia</i>
		<i>Hansenia litangense</i> (0.989 4, 2)	
		<i>Hansenia forrestii</i> (0.988 5, 4)	
14	<i>Heracleum candicans</i>	<i>Heracleum candicans</i> var. <i>candicans</i> (0.995 7, 5)	<i>Heracleum</i>
		<i>Semenovia malcolmii</i> (0.964 3, 1)	
		<i>Similisinocarum pimpinellisimulacrum</i> (0.964 2, 1)	
15	<i>Hymenidium davidii</i>	<i>Hymenidium davidii</i> (1.0, 2)	Apiaceae
		<i>Pleurospermum franchetianum</i> (1.0, 6)	
		<i>Pleurospermum cristatum</i> (0.998 6, 4)	
16	<i>Ligusticum jeholense</i>	<i>Ligusticum jeholense</i> (1.0, 12)	Apiaceae
		<i>Conioselinum anthriscoides</i> (1.0, 8)	
		<i>Conioselinum tenuissimum</i> (1.0, 2)	
17	<i>Nothosmyrnum japonicum</i>	<i>Nothosmyrnum japonicum</i> (1.0, 6)	Apiaceae
		<i>Nothosmyrnum xizangense</i> (0.980 6, 3)	
		<i>Trachyspermum podlechii</i> (0.972 6, 3)	
18	<i>Oenanthe javanica</i>	<i>Oenanthe javanica</i> (1.0, 11)	<i>Oenanthe linearis</i>
		<i>Oenanthe</i> sp. JC-2015 (1.0, 51)	
		<i>Oenanthe sarmentosa</i> (0.986 4, 3)	
19	<i>Oenanthe linearis</i>	<i>Oenanthe benghalensis</i> (0.994 0, 1)	<i>Oenanthe linearis</i>
		<i>Oenanthe javanica</i> subsp. <i>javanica</i> (0.989 9, 13)	
		<i>Oenanthe linearis</i> subsp. <i>linearis</i> (0.985 9, 2)	
20	<i>Ostericum grosseserratum</i>	<i>Ostericum grosseserratum</i> (0.995 7, 7)	Apiaceae
		<i>Ostericum maximowiczii</i> (0.981 2, 10)	
		<i>Angelica anomala</i> (0.976 9, 14)	
21	<i>Pastinaca sativa</i>	<i>Pastinaca sativa</i> (1.0, 6)	<i>Pastinaca sativa</i>
		embryophyte environmental sample (0.998 1, 1)	
		<i>Pastinaca lucida</i> (0.996 0, 1)	
22	<i>Peucedanum japonicum</i>	<i>Peucedanum japonicum</i> (1.0, 9)	<i>Peucedanum</i>
		<i>Peucedanum dissolutum</i> (0.990 5, 1)	
		<i>Peucedanum medicum</i> (0.989 5, 1)	
23	<i>Pimpinella diversifolia</i>	<i>Pimpinella henryi</i> (0.998 6, 1)	<i>Pimpinella</i>
		<i>Pimpinella diversifolia</i> (0.997 9, 7)	
		<i>Pimpinella valleculosa</i> (0.994 5, 1)	
24	<i>Pternopetalum trichomanifolium</i>	<i>Pternopetalum trichomanifolium</i> (1.0, 3)	Apiaceae
		<i>Pternopetalum monophyllum</i> (0.981 0, 1)	
		<i>Pternopetalum vulgare</i> (0.979 1, 6)	
25	<i>Pternopetalum vulgare</i>	<i>Pternopetalum vulgare</i> (1.0, 6)	<i>Pternopetalum</i>
		<i>Pternopetalum trichomanifolium</i> (0.973 9, 3)	
		<i>Pternopetalum monophyllum</i> (0.973 8, 1)	
26	<i>Sanicula chinensis</i>	<i>Sanicula chinensis</i> (0.993 0, 1)	<i>Sanicula</i>
		<i>Sanicula giraldii</i> (0.989 6, 1)	
		<i>Sanicula europaea</i> (0.986 2, 1)	
27	<i>Saposhnikovia divaricata</i>	<i>Saposhnikovia divaricata</i> (1.0, 3)	<i>Saposhnikovia divaricata</i>
		<i>Ledebouriella seseloides</i> (1.0, 2)	
		<i>Cnidium dauricum</i> (1.0, 1)	
28	<i>Torilis scabra</i>	<i>Torilis japonica</i> (1.0, 15)	<i>Torilis</i>
		<i>Torilis africana</i> (0.956 8, 1)	
		<i>Torilis scabra</i> (0.954 1, 1)	

注: 正确的分子鉴定结果用黑体表示。Phylox 鉴定结果括号内的两个数值, 分别代表本地 BLAST 比对的最佳序列一致度(Identity)以及该物种在比对得分前 100 位的候选序列中出现的绝对频次。

Note: The correct molecular identification results are displayed in bold. For Phylox identification results, the two values within the parentheses represent the best sequence identity (Identity) from local BLAST alignment and the absolute frequency of the species among the top 100 candidate sequences in the alignment score.

2.2 三代测序长读段单标记与混合标记分离鉴定结果

本研究进一步选取 8 个三代测序样本,其中样本 1~3 为已预先分离为单标记输入的混合标记样本、4~8 为未分离混合标记样本,所有样本均包含 ITS 与 *psbA-trnH* 两个标记,对升级后的三代模块(UNOISE3 zOTU 路线)进行测试验证,结果见表 2。测试集涵盖了山茶科、无患子科、木犀科、小檗科、马鞭草科、葫芦科及蔷薇科等具有代表性的被子植物类群。

针对 3 个单标记输入样本,系统表现出极高的

捕获效率,经下游 BLAST 比对,所有样本的鉴定结论均与先验信息高度吻合,且生成的 zOTU 序列具有良好的碱基质量。对于 5 个同时包含 ITS 与 *psbA-trnH* 标记的混合标记样本,系统成功实现基于竞争性比对的自动化分流。结果显示,混合标记样本中的两类靶基因均被完整提取,且 BLAST 种级鉴定结果与样本形态学分类信息完全一致。该结果表明,基于竞争性分流与去噪聚类(zOTU)的策略,能够有效克服三代长读段数据的随机测序噪声,为复杂测序情境下植物多条形码联合鉴定提供了高效、稳定的工程化解决方案。

表 2 三代测试集鉴定结果

Tab. 2 Identification results of the third-generation sequencing test dataset

样本编号	形态鉴定结果	ITS 鉴定结果	<i>psbA-trnH</i> 鉴定结果
1	<i>Camellia uraku</i>	<i>Camellia uraku</i> (0.990 9, 1)	<i>Camellia uraku</i> (1.0, 1)
		<i>Camellia japonica</i> (0.989 4, 32)	<i>Camellia japonica</i> (1.0, 6)
		<i>Camellia azalea</i> (0.971 7, 1)	<i>Camellia polyodonta</i> (1.0, 2)
		<i>Koelreuteria bipinnata</i> (0.991 6, 3)	<i>Koelreuteria bipinnata</i> (1.0, 7)
2	<i>Koelreuteria bipinnata</i>	<i>Koelreuteria paniculata</i> (0.990 7, 9)	<i>Koelreuteria elegans</i> (0.951 6, 1)
		<i>Koelreuteria</i> sp. Harder 5724 (0.985 0, 1)	<i>Koelreuteria paniculata</i> (0.949 5, 1)
		<i>Ligustrum</i> × <i>vicaryi</i> (0.998 5, 1)	<i>Ligustrum ovalifolium</i> (1.0, 9)
3	<i>Ligustrum</i> × <i>vicaryi</i>	<i>Ligustrum ovalifolium</i> (0.998 4, 8)	<i>Ligustrum</i> × <i>vicaryi</i> (1.0, 1)
		<i>Ligustrum tschonoskii</i> (0.996 8, 3)	<i>Ligustrum foliosum</i> (1.0, 3)
		<i>Nandina domestica</i> (0.987 0, 4)	<i>Nandina domestica</i> (1.0, 7)
		<i>Caulophyllum robustum</i> (0.868 1, 2)	<i>Aristolochia mollissima</i> (0.948 7, 1)
4	<i>Nandina domestica</i>	<i>Thalictrum minus</i> (0.853 7, 1)	<i>Aristolochia versicolor</i> (0.948 7, 1)
		<i>Vitex negundo</i> (1.0, 18)	<i>Vitex tripinnata</i> (1.0, 1)
		<i>Vitex sampsonii</i> (0.997 2, 1)	<i>Vitex agnus-castus</i> (1.0, 3)
		<i>Vitex agnus-castus</i> (0.994 0, 7)	<i>Vitex negundo</i> (1.0, 16)
5	<i>Vitex negundo</i>	<i>Zehneria bodinieri</i> (0.998 8, 3)	<i>Zehneria monocarpa</i> (0.996 4, 1)
		<i>Zehneria guamensis</i> (0.996 4, 1)	<i>Zehneria tuberifera</i> (0.992 7, 1)
		<i>Zehneria monocarpa</i> (0.992 4, 1)	<i>Zehneria grandibracteata</i> (0.992 7, 1)
		<i>Prunus</i> × <i>dasycarpa</i> (1.0, 2)	<i>Prunus armeniaca</i> (0.983 3, 12)
6	<i>Zehneria monocarpa</i>	<i>Prunus armeniaca</i> (1.0, 82)	<i>Prunus mume</i> (0.978 8, 81)
		<i>Prunus sibirica</i> (0.998 6, 5)	<i>Prunus mandshurica</i> (0.978 8, 2)
		<i>Pyrus uyematsuana</i> (0.989 2, 4)	<i>Pyrus pyrifolia</i> (0.987 8, 5)
		<i>Pyrus calleryana</i> (0.989 2, 38)	<i>Pyrus uyematsuana</i> (0.982 4, 21)
7	<i>Prunus armeniaca</i>	<i>Pyrus pyrifolia</i> (0.986 5, 4)	<i>Pyrus calleryana</i> (0.978 4, 5)

注:正确的分子鉴定结果用黑体表示。Phylox 鉴定结果括号内的两个数值,分别代表本地 BLAST 比对的最佳序列一致性(Identity)以及该物种在比对得分前 100 位的候选序列中出现的绝对频次。

Note:Correct molecular identification results are indicated in bold. The two values in parentheses within the Phylox identification results represent the highest sequence identity (Identity) from local BLAST alignment and the absolute frequency of the species among the top 100 candidate sequences in the alignment score.

2.3 模拟数据基准测试与同类软件对比分析

为系统评估提取与鉴定精度,本节首先在 1.2.4 节所述的金标准模拟数据集上进行定量基准测试,再与主流工具进行对比。在模拟二代扩

增子数据(5 个样本×2 个标记=10 个样本-标记对)上,Phylox 全部 10 对恢复目标条形码(恢复率 100%,平均序列一致性 99.75%),而 Phyluce (SPAdes^[33] 组装+probe 比对)恢复 9 对(恢复率

90%, 平均一致性 99.77%); 在参考分歧较大的位点(如 *Pyrus* ITS), Phylox 仍保持高一致性的局部恢复。在模拟三代 ONT 数据(约 8% 错误)上, Phylox 5/5 正确鉴定到种, UNOISE3 将 zOTU 错误率降至 2%~5%; 5 个真实 ONT 样本亦全部正确鉴定到种。上述模拟数据均源自 1.2.4 节所述的金标准生成流程(二代 wgsim、三代 homopolymer 集中错误), 验证了优化算法在已知真实序列下的提取与鉴定精度。在与主流软件的对比中, Phylox 与 varKoder 在 28 个伞形科二代测序样本上的分类性能与计算时效对比见图 3: Phylox 实现 100% 属级与 92.86% 种级准确率, 明显优于 varKoder 的 64.29% 属级与 28.57% 种级准确率; 统计学检验(t-test)表明两者执行时间差异极显著 $P=0.0002$, 图 3b), Phylox 单样本分析周期控制在 3 min 以内, 优于 varKoder 约 10 min 的平均耗时。需要说明的是, Phyluce 与 PhyProbe^[34] 均为面向二代靶向捕获数据的流程, 不支持三代长读段数据(表 3 中三代部分标注“不支持”)。

在系统功能与数据兼容性层面, 本研究进一步将 Phylox 与经典比对工具 BLAST+、系统发育

微调工具 PhyloTune、扩增子分析管线 QIIME 2 以及条形码分析软件 Dnabarcoder 进行了特征横向对比(表 3)。BLAST+ 和 Dnabarcoder 作为单一功能工具, 严格依赖前置流程提供组装好的纯净 FASTA 序列, 下游仍需人工判定分类阶元; PhyloTune 仅接收已对齐的精简分子标记序列, 用于优化系统发育树拓扑结构; QIIME 2 虽然支持高通量分析, 但其数据源被严格限制在二代扩增子数据中, 且在生态盲测中的分类边界多停留在属/科级(属级准确率约 75%~85%)。相比之下, Phylox 实现了对二代原始短读段与三代原始长读段数据的协同兼容, 采用二代分层迭代提取与三代竞争比对去噪重构的双轨差异化机制, 在原始数据的自动化处理和多标记分流程效能上明显优于传统比对及单一功能条形码工具, 可一键完成从原始数据输入到具备科属交叉验证的物种鉴定报告输出的全流程闭环; 此外, 相比仅提供命令行或单一功能脚本的工具, Phylox 提供的跨平台图形用户界面显著降低了非专业用户的使用门槛, 并内建数据可视化与报告导出功能, 在精细分类阶元的分辨率上亦显著优于新型深度学习模型。

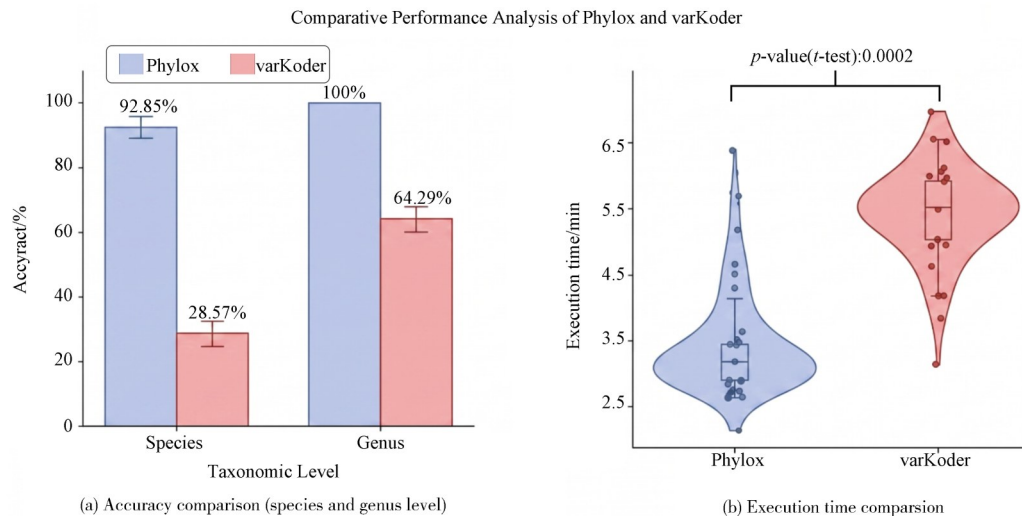


图 3 Phylox 与 varKoder 的性能对比分析

Fig. 3 Performance comparison analysis between Phylox and varKoder

3 讨论

3.1 差异化算法机制与去标签盲测效能验证

分子标记鉴定是植物分类和资源调查的常用手段,但在处理伞形科等形态相似、种间差异小的困难类群时,传统流程在自动化程度和数据通量上存在明显局限。本研究开发的 Phylox 软件通过

整合二代和三代测序数据,实现了从原始数据输入到物种鉴定报告输出的一键式分析。

在数据处理策略上,软件针对不同测序数据的特点采取了差异化算法:对于二代全基因组浅层测序数据,由于测序深度较低且基因组存在高重复区,传统组装容易导致靶向基因片段断裂。本软件通过“科级预筛选”和“属种级精细提取”的

表 3 Phylox 与常见分类学鉴定软件的功能特征横向对比

Tab. 3 Horizontal comparison of functional characteristics and performance metrics between Phylox and conventional taxonomic identification software

软件	输入数据	输出	参考依赖	二代支持	三代支持
Phylox	二代原始短读、三代原始长读	物种鉴定报告(种级+科属交叉验证)	分层本地库(216科)	支持	支持
Phyluce	二代靶向捕获短读	UCE/位点序列与矩阵	需探针/参考	支持	不支持
PhyProbe	二代 WGS/RNA-seq/捕获	目标位点序列	需参考/目标位点	支持	不支持
BLAST+	已组装纯净序列	比对打分列表(需人工判定)	需参考库	部分支持	部分支持
varKoder	二代原始短读	分类标签(科/属/种)	预训练模型	支持	不支持
QIIME 2	二代扩增子	OTU/ASV 表(多为属/科级)	分类器	支持	不支持

两步迭代,并依据拷贝数与倍性自动选择 standard、single_copy 或 phased 组装模式,利用保守的侧翼序列与 UCE 多拷贝守卫定位靶向基因,克服了传统 PCR 扩增偏好、旁系同源误组装与组装碎片化的难题;对于三代长读段数据,针对其单分子测序随机报错率较高的缺陷,软件采用 UNOISE3 去噪直接输出观测真实序列(zOTU),避免共识投票对技术错误的放大,并辅以属级一致性校验剔除污染与嵌合,无需额外引入高精度的二代数据即可独立完成三代读段的校正与多标记(ITS 和 *psbA-trnH*)的自动分流,降低了对计算资源的依赖,也为三代便携式测序仪(如 MinION)在野外现场的快速鉴定提供了可能。

针对困难类群鉴定中可能存在的“循环论证”和标签污染问题,我们在测试中采用了去标签的“盲测”方式,将分析过程与预设的物种标签彻底隔绝,并以权威文献中经过形态与系统发育双重验证的数据作为参考标准。在伞形科独立测试集中,软件实现了 100% 的属级和 92.86% 的种级鉴定准确率,证明该流程在面对近缘种复合体时具备可靠的识别能力。与依靠专家经验且难以处理大批量样本的传统形态鉴定相比,Phylox 将单样本的分析时间缩短到了 3 min 以内,大幅提高了数据分析效率,降低了大批量样本分类鉴定的时间成本。其内置的本地分层数据库覆盖了被子植物 216 个科,涵盖了菊科、兰科、豆科、禾本科等分类学上公认的困难大科。用户只需输入原始测序数据,无需根据不同植物类群调整参数或重写分析流程,即可获得标准化鉴定结果;同时,其模块化架构也允许用户通过替换精简库或区域全库,适配边境检疫或区域普查等特定需求,兼具了标准化分析的稳定性和针对特定任务的可延展性。

3.2 与主流生信工具对比及参考数据库拓补质控

在与现有的生信分析工具对比中,Phylox 展示出明确的实用优势。传统的比对工具过于依赖局部相似度,在面对种间变异极小的近缘物种时容易出现多重歧义匹配;基于宏观特征或深层神经网络的深度学习模型虽然在科/属等高阶元上表现较好,但在缺乏特征位点权重的种级识别上容易模糊边界;扩增子分析管线则主要针对群落多样性分析,缺少对二代与三代异质数据的一键式流转支持。Phylox 通过将靶向基因提取与子库建树校验结合,在保障种级识别精度的同时兼顾了分析效率。测试中少数样本仅能鉴定到属,主要原因是极端近缘种之间的序列过于相似。这说明当前参考数据库在困难类群的变异位点覆盖上仍有不足,补充高质量的全长标准参考序列是进一步提升种级识别率的关键。更重要的是,Phylox 通过多分辨率参考检索与低参考依赖策略,在缺乏近缘参考的非模式物种场景中仍能工作,而其他以高完备参考/探针为前提的工具在此类场景中恢复率受限——这正是 Phylox 独有适配的价值所在。

Phylox 的价值不仅在于提高鉴定效率,更在于通过自动化流程缓解全球分类学专业人才萎缩(即“分类学障碍”)带来的工作积压,帮助分类学家将精力集中于少数疑难样本与新物种描述。更为重要的是,该软件内置的拓扑冲突检测机制可在鉴定过程中自动识别参考数据库中可能存在错误标注或分类归属不当的序列。在伞形科测试中,我们即发现 GenBank 中部分 ITS 序列存在属级误标和种名张冠李戴的问题,这些错误序列若不经质控直接纳入分析,将对鉴定结果和下游系统发育推断产生误导。Phylox 的自动清理功能实质上是对参考数据库的持续“纠错”,伴随分析样

本的积累,这种纠错效应将逐步提升公共核酸数据库的整体可靠性。此外,软件输出带有置信度标注的系统发育树,使得每个鉴定结果都可追溯至具体的拓扑证据,兼顾了分类鉴定结果的准确性与可检验性。

3.3 鉴定误差的生物学归因与应用展望

客观来看,系统在处理个别复杂样本时出现的错鉴或无法赋值,主要源于生物学特性和参考数据库限制两个方面。在生物学层面,经历过近期辐射演化、种间杂交或存在严重不完全谱系分流(ILS)的近缘类群,常规条形码标记的种间遗传距离可能小于种内变异,导致单基因序列难以提供足够的分类信息;在数据库层面,公共数据库中普遍存在的误标和序列污染会严重干扰自动化鉴定,本软件虽然加入了基于系统发育拓扑冲突的动态质控算法,但在面对缺乏高质量参考序列的困难类群时,为了保证结果的可靠性,系统会采取保守策略而放弃赋值。在模拟二代与三代数据集上的定量基准测试中,Phylox 在恢复率上更具优势,序列质量与主流工具相当。

综上,本软件在数据兼容性与自动化闭环分析方面的实用性得到了验证。在实际应用方面,该软件在中药材基原鉴定、植物资源普查和生物多样性监测中具有良好的工程应用价值,其对三代混合标记样本的成功分流也验证了对复杂或降解样本的适用性,可为植物分子鉴定与资源保护提供稳定可靠的工具支持。

后续优化将从三个方面展开:一是扩展条形码种类,发挥软件模块化架构的优势,吸收叶绿体全基因组和 rDNA 完整转录单元等更大数据集,通过多标记联合鉴定解决单基因片段信息量不足的问题,并可拓展至动物与微生物数据的快速分类;二是升级底层算法,引入轻量化机器学习算法,提高对低覆盖度或严重降解样本的容错能力;三是拓展应用场景,在已提供跨平台图形交互界面(Tauri)的基础上,进一步拓展长读长(HiFi/ONT)支持、环境 DNA/宏条形码模式与自动化基准测试框架,并保持参考数据库的持续更新维护。数据与代码可用性声明:

Phylox 的安装指南、详细文档、运行脚本、二代与三代测试数据集及完整源代码均已开源托管于 Gitee 平台(<https://gitee.com/sculab/phylox>),并提供命令行(phyx)与跨平台桌面图形界面两种

使用入口,支持 Linux、Windows 与 macOS 操作系统。基准测试的模拟数据生成器、评估脚本与中间结果已一并开源,保证定量对比可复现。

参考文献:

- [1] Group C P W, Hollingsworth P M, Forrest L L, *et al.* A DNA barcode for land plants[J]. *Proc Natl Acad Sci U S A*, 2009, 106(31): 12794-12797.
- [2] Hebert P D N, Cywinska A, Ball S L, *et al.* Biological identifications through DNA barcodes [J]. *Proc R Soc Lond B*, 2003, 270(1512): 313-321.
- [3] Kress W J, Erickson D L. DNA barcodes: Genes, genomics, and bioinformatics [J]. *Proc Natl Acad Sci U S A*, 2008, 105(8): 2761-2762.
- [4] Kress W J. Plant DNA barcodes: Applications today and in the future [J]. *J Syst Evol*, 2017, 55(4): 291-307.
- [5] DeSalle R, Goldstein P. Review and interpretation of trends in DNA barcoding [J]. *Front Ecol Evol*, 2019, 7: 302.
- [6] Zhu S, Liu Q, Qiu S, *et al.* DNA barcoding: an efficient technology to authenticate plant species of traditional Chinese medicine and recent advances [J]. *Chin Med*, 2022, 17: 112.
- [7] Group C P B, Li D Z, Gao L M, *et al.* Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants [J]. *Proc Natl Acad Sci U S A*, 2011, 108(49): 19641-19646.
- [8] Pang X, Liu C, Shi L, *et al.* Utility of the trnH-psbA intergenic spacer region and its combinations as plant DNA barcodes: a meta-analysis [J]. *PLoS One*, 2012, 7(11): e48833.
- [9] Bolson M, de Camargo Smidt E, Brotto M L, *et al.* ITS and trnH-psbA as efficient DNA barcodes to identify threatened commercial woody angiosperms from southern Brazilian Atlantic rainforests [J]. *PLoS One*, 2015, 10(12): e0143049.
- [10] Frigerio J, Agostinetto G, Mezzasalma V, *et al.* DNA-based herbal teas' authentication: an ITS2 and psbA-trnH multi-marker DNA metabarcoding approach [J]. *Plants*, 2021, 10(10): 2120.
- [11] Tripathi A M, Tyagi A, Kumar A, *et al.* The internal transcribed spacer (ITS) region and trnH-psbA are suitable candidate loci for DNA barcoding of tropical tree species of India [J]. *PLoS One*, 2013, 8(2): e57934.
- [12] Goodwin S, McPherson J D, McCombie W R.

- Coming of age: ten years of next-generation sequencing technologies[J]. *Nat Rev Genet*, 2016, 17(6): 333-351.
- [13] Batovska J, Cogan N O I, Lynch S E, *et al*. Using next-generation sequencing for DNA barcoding: capturing allelic variation in *ITS2* [J]. *G3 Genes|genomes|genetics*, 2017, 7(1): 19-29.
- [14] Hollingsworth P M, Li D Z, van der Bank M, *et al*. Telling plant species apart with DNA: from barcodes to genomes [J]. *Phil Trans R Soc B*, 2016, 371(1702): 20150338.
- [15] Coissac E, Hollingsworth P M, Lavergne S, *et al*. From barcodes to genomes: extending the concept of DNA barcoding[J]. *Mol Ecol*, 2016, 25(7): 1423-1428.
- [16] Tonti-Filippini J, Nevill P G, Dixon K, *et al*. What can we do with 1000 plastid genomes [J]. *Plant J*, 2017, 90(4): 808-818.
- [17] Murigneux V, Rai S K, *et al*. Comparison of long-read methods for sequencing and assembly of a plant genome [J]. *GigaScience*, 2020, 9(12): giaa146.
- [18] Jiao W B, Schneeberger K. The impact of third generation genomic technologies on plant genome assembly [J]. *Curr Opin Plant Biol*, 2017, 36: 64-70.
- [19] Gostel M R, Zúniga J D, Kress W J, *et al*. Microfluidic Enrichment Barcoding (MEBarcoding): a new method for high throughput plant DNA barcoding [J]. *Sci Rep*, 2020, 10: 8701.
- [20] Letsiou S, Madesis P, Vasdekis E, *et al*. DNA barcoding as a plant identification method [J]. *Appl Sci*, 2024, 14(4): 1415.
- [21] Deng D, Xu W, Wu B, *et al*. PhyloTune: an efficient method to accelerate phylogenetic updates using a pretrained DNA language model [J]. *Nat Commun*, 2025, 16: 6905.
- [22] de Medeiros B A S, Cai L, Flynn P J, *et al*. A composite universal DNA signature for the tree of life [J]. *Nat Ecol Evol*, 2025, 9(8): 1426-1440.
- [23] Raclariu A C, Heinrich M, Ichim M C, *et al*. Benefits and limitations of DNA barcoding and metabarcoding in herbal product authentication [J]. *Phytochem Anal*, 2018, 29(2): 123-128.
- [24] Vu D, Nilsson R H, Verkley G J M. Dnabarcoder: An open-source software package for analysing DNA barcoding data [J]. *Molecular Ecology Resources*, 2022, 22(8): 2793-2804.
- [25] Banchi E, Ametrano C G, Greco S, *et al*. PLANITS: A curated sequence reference dataset for plant ITS DNA metabarcoding [J]. *Database*, 2020, 2020: baz155.
- [26] Oliveira R R M, Nunes G L, de Lima T G L, *et al*. PIPEBAR and OverlapPER: Tools for a fast and accurate DNA barcoding analysis and paired-end assembly [J]. *BMC Bioinform*, 2018, 19: 297.
- [27] Yu X, Tang Z, Zhang Z, *et al*. GeneMiner2: Accurate and automated recovery of genes from genome skimming data[J]. *Mol Ecol Resour*, 2026, 26(2): e70111.
- [28] Dormontt E E, van Dijk K-J, Bell K L, *et al*. Advancing DNA barcoding and metabarcoding applications for plants requires systematic analysis of herbarium collections—An Australian perspective [J]. *Front Ecol Evol*, 2018, 6: 134.
- [29] Kolter A, Hebert P. Automating the curation of DNA Barcode databases for vascular plants[J]. *Environ DNA*, 2025, 7(3): e70125.
- [30] Edgar R C, Flyvbjerg H. Error filtering, pair assembly and error correction for next-generation sequencing reads[J]. *Bioinformatics*, 2015, 31(21): 3476-3482.
- [31] Edgar R. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing[EB/OL]. [2026-05-31]. <https://www.biorxiv.org/content/10.1101/081257v1>.
- [32] Faircloth B C, McCormack J E, Crawford N G, *et al*. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales [J]. *Syst Biol*, 2012, 61(5): 717-726.
- [33] Bankevich A, Nurk S, *et al*. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing [J]. *J Comput Biol*, 2012, 19(5): 455-477.
- [34] Rautsaw R M. PhyProbe: A bioinformatic pipeline to extract phylogenetic loci from next-generation sequencing data [EB/OL]. [2026-05-31]. <https://github.com/RhettRautsaw/PhyProbe>.

(责任编辑:白林含)